Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–17

DOI: 10.1177/0962280215620696
smm.sagepub.com
SAGE

Controlled multi-arm platform design using predictive probability

Brian P Hobbs, Nan Chen and J Jack Lee

Abstract

The process of screening agents one-at-a-time under the current clinical trials system suffers from several deficiencies that could be addressed in order to extend financial and patient resources. In this article, we introduce a statistical framework for designing and conducting randomized multi-arm screening platforms with binary endpoints using Bayesian modeling. In essence, the proposed platform design consolidates inter-study control arms, enables investigators to assign more new patients to novel therapies, and accommodates mid-trial modifications to the study arms that allow both dropping poorly performing agents as well as incorporating new candidate agents. When compared to sequentially conducted randomized two-arm trials, screening platform designs have the potential to yield considerable reductions in cost, alleviate the bottleneck between phase I and II, eliminate bias stemming from inter-trial heterogeneity, and control for multiplicity over a sequence of a priori planned studies. When screening five experimental agents, our results suggest that platform designs have the potential to reduce the mean total sample size by as much as 40% and boost the mean overall response rate by as much as 15%. We explain how to design and conduct platform designs to achieve the aforementioned aims and preserve desirable frequentist properties for the treatment comparisons. In addition, we demonstrate how to conduct a platform design using look-up tables that can be generated in advance of the study. The gains in efficiency facilitated by platform design could prove to be consequential in oncologic settings, wherein trials often lack a proper control, and drug development suffers from low enrollment, long inter-trial latency periods, and an unacceptably high rate of failure in phase III.

Keywords

Bayesian analysis, multi-arm controlled clinical trial design, multiple comparisons, predictive probability, sequential design

1 Introduction

The pursuit of personalized medicine has accelerated the pace of scientific discovery in the fields of molecular biology, genomics, proteomics, and metabolomics, etc. producing enormous numbers of

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Corresponding author:
Brian P Hobbs, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA.
Email: bphobbs@mdanderson.org

molecules, yielding numerous avenues for the development of potentially effective drug therapies. Yet, the system for clinical testing continues to suffer from inherent inefficiencies. The current system for drug development was devised to screen and test experimental agents one-at-a-time in phases (I, II, and III), over the course of a sequence of clinical trials, each designed to evaluate different types of endpoints. Often each trial requires an enrollment period spanning multiple years to attain the targeted sample size. Thereafter, a follow-up period is required to ascertain the extent of therapeutic response for all patients prior to statistical analysis. Moreover, each transition between phases involves a latency period wherein the next study is designed and reviewed prior to initiation. In oncology, these latency periods (sometimes referred to as operational "whitespace") span a duration of nearly 2 years on average.[1] In fact, the authors of a recent National Cancer Institute bulletin[2] surmise that evaluation of every possible combination of only 10 investigational drug regimens under the current system would require approximately 90 years to complete. Moreover, this system for screening agents one-at-a-time also suffers from low specificity, failing to identify ineffective therapies (true negatives) for which further development should be stopped. This is evident from the limited recent success in oncology wherein only 34% of confirmatory phase III trials yielded a significant result from 2003 to 2010[3] and final market approval was achieved for only 13% of the cancer drugs that initiated phase I between 1993 and 2004.[4]

There is growing consensus that revision is necessary if the system is to accommodate efficient clinical testing of the numerous emerging therapies formulated to target specific molecular pathways. The Pharmaceutical Research and Manufacturers of America reports that currently 771 new cancer medicines and vaccines developed by U.S.-based biopharmaceutical companies have initiated clinical trials or await regulatory review.[5] Moreover, pre-clinical evaluations are currently underway for thousands of compounds.[6] A recent report on the cooperative oncology groups by the Institute of Medicine[1] even goes as far as advocating for the need to restructure the entire clinical trials system for the purpose of avoiding redundancy and improving effectiveness and efficiency. The Clinical Trial Design Task Force (CTD-TF) of the National Cancer Institute's (NCI) Investigational Drug Steering Committee recently emphasized several types of design modifications for prospective trials[7] that could be implemented to enhance efficiency without sacrificing statistical integrity. These include sequential learning from accumulating trial data for early termination, possible extension, and to establish or identify predictive subgroups; multi-stage designs that enable seamless phase transitions; and designs that allow mid-trial incorporation of new arms.

We contend that in many areas of medicine, the aforementioned objectives may be more easily attainable by reforming phase II itself from a system consisting of many disparate, sequentially conducted, weakly informative, often non-comparative trials into a system that uses sequentially adaptive, controlled multi-arm screening platforms designed to test competing therapies simultaneously as they emerge. In this article, we introduce a statistical framework for designing and conducting randomized multi-arm screening platforms using Bayesian modeling. The platform approach necessitates mid-trial modifications to the study arms that enable both dropping poorly performing agents as well as adding new agents at random entry times immediately after a proper dose has been selected in phase I. Thus, the environment yields imbalanced comparisons. The primary statistical issue to resolve in the platform setting pertains to implementation of sequentially adaptive futility monitoring while controlling familywise type I error at a pre-specified level when the number of total comparisons is itself stochastic.

The proposed platform-based approach to phase II offers several advantages. Consolidating inter-study control arms to form a concurrent control enables investigators to assign more new patients to novel therapies, thereby screening more agents in a shorter time span. Facilitating contiguous platforms for incorporating new agents as they emerge, via a protocol amendment to

an ongoing trial in place of the processes for design, review, and approval of a new study, promotes systemic efficiency by reducing latency periods between phase I and II. Enabling investigators to compare the clinical effectiveness of experimental agents formulated to target different signaling pathways or mechanisms for regulating response to a common standard of care therapy would eliminate bias stemming from inter-trial heterogeneity and better control for multiplicity, thereby improving the quality of the primary drug development decision regarding whether to proceed to phase III. Thus, the approach promises to better avoid expensive confirmatory phase III studies for ineffective agents. The concept of platform design had been discussed by Lee and Chu.[8] In addition, recently several authors have also explored the extent to which multi-arm[9–11] and multi-stage[12,13] randomized designs may yield improvements in efficiency and enhance treatment efficacy.

We build the framework for platform design in the following sequence. Section 2 reviews approaches for phase II design. Section 3 introduces the concept of platform-based screening. Section 4 presents the Bayesian probability model for inference. Section 5 introduces the appropriate continuous futility monitoring rule derived from predictive probability (PP). Section 6 presents both design considerations, as well as the results of our investigation of the operating characteristics for platform design. Here we demonstrate that platform-based approaches offer considerable gains in efficiency when matched to attain identical frequentist properties for inference under the conventional one-at-a-time paradigm using a sequence of sequential two-arm trials. Section 7 discusses trial conduct under the platform approach, which can be implemented readily without advanced statistical software. Finally, discussion and concluding remarks are provided in Section 8.

2 Experimental designs for intermediate-phased clinical testing

2.1 Phase II clinical trial design

The motivation for conducting a phase II clinical trial is to decide whether to pursue more extensive development.[14,15] Traditionally, these go/no-go decisions were based on small non-comparative studies designed only to estimate the extent to which the agent under study demonstrates sufficient biological activity to induce therapeutic response. Response is typically defined to be a binary endpoint that is measured after a short duration following administration of the therapy. Conventionally, early phase II trials are single-arm studies, with measures of evidence for effect based on ''comparison'' to results from a historical cohort or fixed null response rate.

The traditional phase II study for cancer research was a two-stage design proposed by Gehan[14] wherein continuation criteria was evaluated at a single interim analysis. The design facilitates early termination in the absence of early evidence for efficacy. Thereafter, two-stage and multistage designs based on the multiple-testing procedure and group sequential theory proposed in literature[15–17] provided sequential adaptivity, enhancing efficiency even further. In addition, Yao et al.[18] proposed a process for screen multiple vaccines in a series of trials formulated to test treatments one-at-a-time until a promising agent is identified.

Thall and Simon[19] provided some practical guidelines on how to implement a phase II trial from the Bayesian paradigm. This design was the first to enable continuous trial monitoring using posterior probability updated after observing each patient response. The decision to stop the trial (on the basis of absence of effect or strong evidence for improvement) or continue enrolling patients (because of a lack of convincing evidence to inform a decision) is reassessed throughout the conduct of the trial until a maximum sample size has been attained. Lee and Liu[20] introduced another Bayesian method for continuous monitoring of a single-arm phase II trial based on the PP of a successful result at the end of the trial given continuation to a pre-specified sample size.

Despite these advances, single-arm phase II trials suffer from inherent limitations.[21] Go/no-go decisions resulting from single-arm trials are vulnerable to inter-trial heterogeneity and neglect to control for selection bias, risking biased comparisons that often over-exaggerate the evidence for future success in phase III. Randomization ensures patient comparability, on average, yielding statistically valid estimates of the treatment effect. Thus, randomization is especially crucial in an environment with limited resources for phase III. In fact, the Clinical Trial Design Task Force of National Cancer Institute suggested that the recent limited success of phase III oncology trials is partially due to the lack of randomized comparisons in phase II. Randomized phase II selection designs, first introduced by Simon et al.,[22] were based on ranking and selection among multiple active treatments as an alternative to testing the null hypothesis of therapeutic equivalence. In phase II settings, wherein the "recruitment of patients does not keep pace with the supply of novel therapies," Whitehead[23] proposed optimal Bayesian criteria for selecting the number of treatments and sample size for each treatment one should evaluate in a series of uncontrolled pilot studies with randomized assignment among treatments that are available simultaneously. More recently, Lee and Feng[24] examined the impact of randomization on phase II design in oncology. Yin et al.[25] extended the PP approach to controlled adaptively randomized phase II design.

2.2 Optimising phase II in consideration of phase III

Typically, sufficient success in phase II leads to continued testing in phase III wherein a much larger randomized trial is implemented to evaluate the extent to which the experimental therapy may prolong time to treatment failure or survival when compared to one or more standard of care therapies. A few authors have proposed statistical approaches for designing a phase II study such that the entire evaluation process (phase II and phase III) might be considered optimal. Whitehead[26] considered the setting wherein phase III, which is assumed to consist of a two-arm controlled randomized study that is designed to compare binary endpoints, is preceded by a series of uncontrolled pilot studies in phase II that are implemented to select the best candidate therapy for phase III. More specifically, assuming that the response rates for candidate therapies are exchangeable and described by a beta distribution, Whitehead derived optimal strategies for jointly specifying the total number of patients required for both phases, the proportion that should be allocated to each phase, and the total number of therapies that should be tested. Considering the setting wherein phase II consists of a sequence of controlled, randomized two-arm trials each possibly proceeded by phase III in the event that the phase II trial yields a statistically significant improvement for the experimental therapy, Stallard[27] established Bayesian decision-theoretic approaches for designing the entire clinical evaluation program to minimize the total number of patients required to conduct a successful phase III trial. Wason et al.[28] extended the approach to accommodate screening trials in phase II that test multiple new therapies simultaneously.

3 Platform design in phase II

As noted above, nearly two out of every three confirmatory oncology trials fail. Therefore, demonstrating definitive efficacy following success in phase II is more likely to fail than succeed under the traditional screen one-drug-at-a-time paradigm. The high failure rate also suggests an overreliance on small, single-arm trials, inadequate surrogate endpoints for long-term efficacy, and lack of proper active controls. While results from phase II are not intended to serve as the basis for

guiding medical practice,[29] the go/no-go decision for phase III is inherently comparative and should derive from a comparative study when possible.

Figure 1 provides a conceptual illustration of two approaches for screening agents in phase II. The horizontal axis represents calendar time, while increases in the direction of the vertical axes represent increasing trial enrollment. The top panel depicts a sequence of randomized, sequential two-arm phase II trials. Here we see a series of paired stacked right triangles. These represent enrollment for each of the five two-arm group sequential studies. Their identical size indicates equal allocation between the two study arms. Smaller triangles reflect earlier termination. Gaps between the triangles represent inter-trial latency periods of inactivity reflecting the "whitespace" or periods required to design, review, and initiate each trial.

The platform design, shown below, represents a continuous screening process for comparing multiple agents simultaneously to a common concurrent control. The version depicted uses

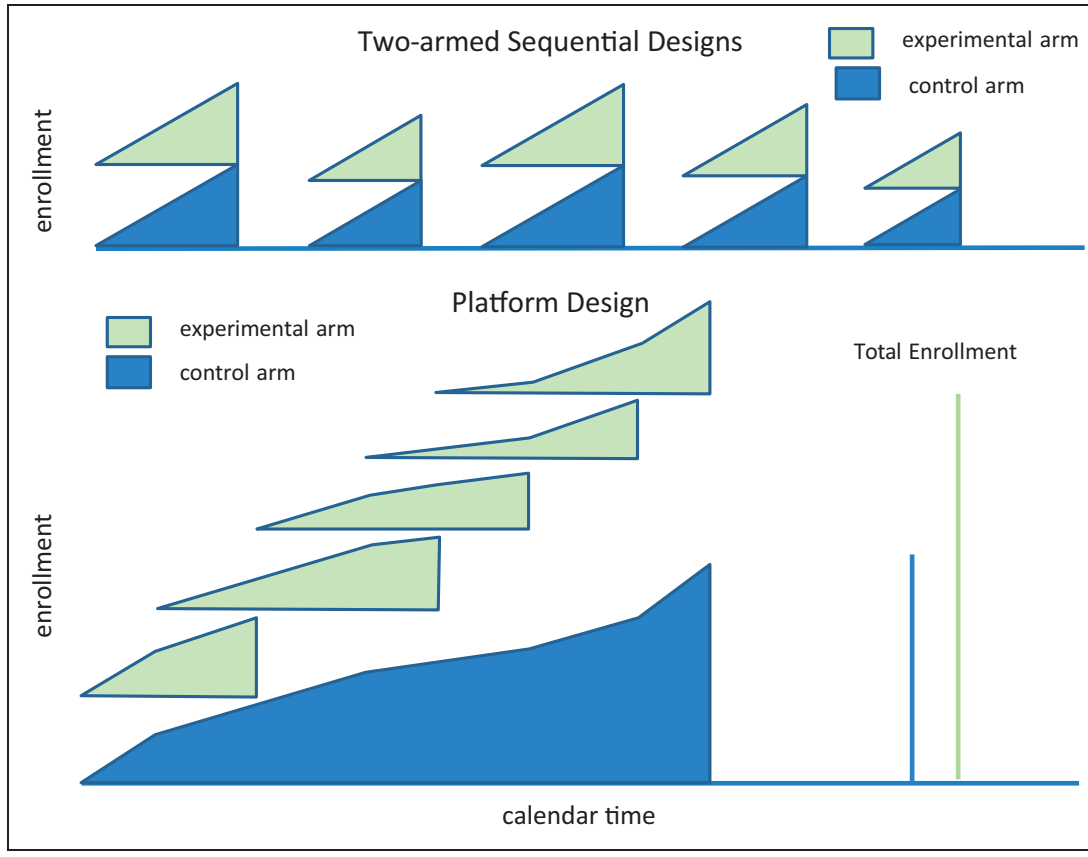

Figure 1. Illustrations of the sequential two-arm and proposed platform-based approaches to randomized phase II design. Each shape depicts a study arm. The horizontal axis represents calendar time. Increases in the direction of the vertical axes represent increasing enrollment. The randomized two-arm approach necessitates that the standard of care therapy is repeated five times. This is reflected in the top panel by five disparate blue-colored right triangles. The platform design enables consolidation of the control arms as well as seamless incorporation of novel investigational agents and as they emerge. This reduces redundancy and enhances efficiency.

randomization to allocate patients equally among all concurrently active arms (represented by identical slopes among active arms at a given trial time). Enrollment to experimental arms is either terminated early due to futility or continues until a pre-specified maximum sample size (per arm) is attained. The approach is similar in nature to a multi-arm selection trial, yet overcomes it's major weaknesses: the absence of an active comparator for uncontrolled selection designs and/or lack of control of type I error.[30] Note that, similar to the traditional phase II design for initial efficacy screening, the platform design does not allow early stopping for superior efficacy. Unlike confirmatory trials in phase III, we often lack the disposition to conclude "superiority" on the basis of a partially completed phase II study. In this setting, agents that are observed to be putatively effective initially would warrant continued enrollment to substantiate the efficacy finding and assess molecular markers of various types to discern trends among genomic-based subsets and determine which biologically-defined patient population is most likely to benefit. Thus, if the agent is working well in comparison to an established standard of care therapy on the basis of a phase II study enrolling only a limited patient population, it is desirable to continue enrolling more patients to learn more about the treatment's efficacy and refining its safety profile.

Comparing the top and bottom panels reveals obvious structural deficiencies underlying the one-at-a-time approach that are overcome by the platform design. The sequence of two-arm studies suffers from inefficient phase I/II transitions. An operational platform enables seamless incorporation of novel investigational agents as they emerge. Screening five arms using the randomized two-arm one-at-a-time approach necessitates that the standard of care therapy is repeated five times in the top panel. The platform design enables consolidation of the control arms, thereby lessening redundancy. As a consequence, investigators assign more new patients to novel therapies, thereby screening more agents within a given time period. In this hypothetical example, the consolidated control arm in the platform enrolls to an extent of about 2/3 of the total enrollment among the five control arms. In addition, the study duration required to screen only 3.15 experimental agents using the two-arm sequential approach is sufficient to complete testing of all five agents under the platform design. Finally, trials for screening agents one-at-a-time fail to acknowledge inherent multiple comparisons over a sequence of a priori planned studies. Thus, this paradigm neglects to control the nominal type I error rate. By way of contrast, the platform design can be calibrated to control familywise type I error which will be demonstrated in subsequent sections.

4 Bayesian probability model

This section presents details of the probability models that will be used to formulate decision rules for monitoring futility during enrollment, and deciding whether sufficient "improvement" is evident given that the maximum sample size is achieved. Phase II trials are designed to evaluate therapeutic activity as measured by the appropriate short-term clinical response. Ideally, the outcome is an established surrogate marker for long-term efficacy. In oncology, the short-term response variable, usually measurable within weeks of administration of therapy, derives almost invariably from dichotomizing an ordinal-valued composite endpoint combining outcomes both local and distant to the treated disease site. For solid tumors, it is common to use either the absence of progressive disease or attainment of a complete response or partial response as defined by RECIST which considers both the extent of change in tumor size on imaging and the presence of new lesions or distant progression within some pre-specified duration from treatment.[31] In some settings such as lung cancer, "disease control" is defined as the absence of progression after 8 weeks. Alternatively, response could be defined by the presence/absence of residual disease as measured by

the extent of malignant cellularity (i.e. <1%) acquired from pathologic evaluation of an excised specimen. This endpoint has been shown to correlate with overall survival.[32–34]

Before presenting details of the model, we need to introduce notation for the observables. Let $Y_{j,i}$ denote the binary random variable indicating whether the ith patient treated with therapy j experienced a successful therapeutic response. Let $j=0$ represent the common standard of care therapy and assume that the platform is designed to incorporate a maximum of J experimental arms. Let $n_j = n_j(t)$ denote the number of patients treated with the jth therapy for which response has been observed by trial time $t > 0$.

We'll use parameter π_j which denotes the probability that therapy j yields a therapeutic response. Assuming the conjugate beta prior distribution, $p(\pi_j) = Beta(a_j, b_j),\ a_j, b_j > 0$, yields an analytically tractable marginal model, from which we can attain an exact expression for the probability of observing a therapeutic response for a future patient given the interim data. The posterior for π_j after observing r_j responses from n_j patients is the following beta distribution

$$f(\pi|r_j) = \frac{1}{\beta(r_j + a_j, n_j - r_j + b_j)} \pi^{r_j + a_j - 1}(1-\pi)^{n_j - r_j + b_j - 1}, \quad (1)$$

where $\beta()$ represents the beta function and $0 \le \pi \le 1$. Let $N_{\max}$ denote the planned maximum number of patients who may enroll into each experiment arm and let $R_j(N_{\max}, \boldsymbol{Y_j}) = R_j$ denote a random variable counting the number of responders among $N_{\max}$ patients treated with the jth therapy. The probability of observing s successes (responses) among $N_{\max} - \mathrm{n_j}$ future patients can be expressed as a product of gamma functions

$$\Pr(R_j = r_j + s|r_j) = \frac{\Gamma(N_{\max} - n_j + 1)}{\Gamma(s+1)\Gamma(N_{\max} - n_j - s + 1)} \frac{\Gamma(s + r_j + a_j)\Gamma(N_{\max} - s - r_j + b_j)}{\Gamma(N_{\max} + a_j + b_j)} \times \frac{\Gamma(n_j + a_j + b_j)}{\Gamma(r_j + a_j)\Gamma(n_j - r_j + b_j)}, \quad (2)$$

where $s = 0, \ldots, N_{max} - n_j$.

5 Futility monitoring based on PP

In the event that interim data from an ongoing phase II trial provide evidence to suggest that the agent is unlikely to meet the pre-specified efficacy objectives, investigators should halt the trial to protect patients as well as re-examine the dose, schedule, targeted patient population, or whether the agent should be re-considered in a future non-inferiority study. Because novel experimental therapies are permitted to enter an ongoing platform mid-trial, two-sample comparisons with the common control are likely to be imbalanced. Conventional methods for futility monitoring (from both frequentist and Bayesian paradigms) base the decision to terminate enrollment on estimates of the model parameters. In the presence of sample size imbalance, these methods require ad-hoc burn-in periods which reduce efficiency by imposing that the trial attains a minimum the number of events before applying the decision rule.

We overcome this issue by defining the decision criterion for futility stopping as a function of the posterior PP of a successful trial. Conceptually, the PP calculation uses the interim data to measure the extent to which investigators should expect a positive result given that enrollment continues up to the targeted sample size, $N_{\max}$. The Bayesian model is used to account for both sources of uncertainty arising from interim estimation of the model parameters (through the posterior), and

the extent of variability in ascertaining the responses of future, heretofore unobserved patients (through the posterior predictive density).

In the context of a randomized screening design, "success" for treatment j implies acquiring sufficient evidence to conclude that the trial demonstrated an improvement in the response rate when compared to the control. Mathematically, the binary decision pertaining to whether sufficient improvement was evident is an evaluation of the posterior probability

$$\Pr(\pi_j > \pi_0 + \delta | R_0, R_j) > \theta.\psi$$

Argument $\delta > 0$ determines the extent to which an improvement is clinically significant, while the posterior threshold $\theta \in (0, 1)$ controls the amount of "evidence" required to conclude success. Using properties of the beta distribution (1), this posterior probability follows as

$$\Pr(\pi_j > \pi_0 + \delta | R_0, R_j) = \int_0^{1-\delta} \Pr(\pi_j > \pi + \delta | R_j) f(\pi | R_0 + a_0, N_{max} - R_0 + b_0) \mathrm{d}\pi \quad (3)$$

where

$$\Pr(\pi_j > \pi + \delta | R_j) = 1 - \frac{\int_0^{\pi+\delta} u^{R_j + a_j - 1}(1-u)^{N_{\max} - R_j + b_j - 1} du}{\beta(R_j + a_j, N_{\max} - R_j + b_j)}.$$

In the presence of incomplete enrollment, n_0, $n_j < N_{\max}$, the PP that the trial ultimately demonstrates improvement for treatment j follows from equations (2) and (3) as

$$\begin{aligned} \lambda(r_o, r_j) &= E_{R_0, R_j}[I\{\Pr(\pi_j > \pi_0 + \delta | R_0, R_j) > \theta\}] \\ &= \sum_{u=0}^{N_{\max} - n_0} \sum_{v=0}^{N_{\max} - n_j} I\{\Pr(\pi_j > \pi_0 + \delta | r_0 + u, r_j + v) > \theta\} \\ &\quad \times \Pr(R_0 = r_0 + u | r_0) \Pr(R_j = r_j + v | r_j) \end{aligned} \quad (4)$$

where $I\{\}$ represents the indicator function. The decision to terminate enrollment to the jth treatment after observing n_j patients follows from evaluating the PP of eventual success

$$\begin{aligned} &\lambda(r_0, r_j) < \phi, \text{ terminate enrollment to the } jth \text{ experimental therapy for futility}, \\ &\lambda(r_0, r_j) \geq \phi, \text{ continue enrolling patients to the } jth \text{ experimental therapy}, \end{aligned} \quad (5)$$

for a given threshold $\phi \in (0, 1)$.

Unlike methods that rely on posterior probability, sequential decisions based on PP account for both uncertainty among information heretofore acquired and the extent of variability for outcomes yet to be observed in the trial. This imparts robustness to sample size imbalance, which is the primary statistical challenge to implementing a screening platform for which investigators decide to maintain the common standard of care as a treatment option for all patients. As new agents become integrated into an ongoing platform, n_0 may already exceed the targeted experimental maximum sample size, $N_{\max}$. In this case, we may compute $\lambda(r_0, r_j)$ conditional on the observed controls, $\sum_{v=0}^{N_{\max} - n_j} I\{\Pr(\pi_j > \pi_0 + \delta | r_0, R_j + v) > \theta\} \Pr(R_j = r_j + v | r_j)$.

The PP approach offers additional advantages. PP-derived rules can be applied uniformly at any time during the trial without the need to specify and calibrate an arbitrary burn-in period or impose

ad-hoc rules for adjusting the decision thresholds in relation to interim sample size (such as those proposed in Wathen and Thall[35]). Moreover, they have been shown to yield rejection regions with smoother transitions when compared with posterior methods and provide higher early stopping probability under null scenarios.[20,25] The following sections demonstrate how PP-derived futility monitoring can be used to design and implement a continuous screening platform that is calibrated to deliver desirable frequentist properties given a pre-specified maximum number of comparisons.

6 Trial design and performance

In this section, we address considerations for using simulation as a tool for calibrating the parameters of the platform designs to attain acceptable frequentist properties. Thereafter, we evaluate the extent to which consolidation can improve efficiency by comparing the resultant induced operating characteristics for platform design to those attained from sequential design using the one-drug-at-a-time evaluation paradigm.

6.1 Simulation approach

Operating characteristics for the screening platform were compared to the corresponding two-arm sequential PP design[25] when applied over the course of five successive trials. Emulating the phase II oncology setting, we assumed a true null response rate of 0.2, set the indifference boundary to $\delta = 0.1$, and assumed that therapeutic response was evaluated after a latency period of 4 weeks following therapy. The Bayesian model for both designs assumed identical $Beta(1, 1)$ priors for all treatment response probabilities, reflecting maximum entropy with prior effective sample size of 2. We assumed that each simulated platform trial was designed to test a maximum of five experimental arms with $N_{\max} = 70$ per arm. Permuted-block randomization with block-size equal to the number of active arms was used for both designs. For tracking calendar time, we assumed an enrollment rate of 10 patients per month. Results were computed from 2000 simulated trials.

6.2 Design considerations

The PP-derived futility calculation (4) requires pre-specification of three arguments. δ determines the interval of clinical indifference or the minimum improvement required to achieve clinical relevance. This must be pre-specified in consideration of several factors including the expected null response rate and other external conditions underlying the trial. For example, the choice of δ could be influenced by the effectiveness of available treatment options for patients with the disease understudy. In an environment where there are many competing agents, one may choose to impose stricter criteria for clinical improvement by increasing the value of δ. The respective posterior and PP thresholds, θ and ϕ, are tuning parameters. These must be calibrated at the design stage to yield desirable operating characteristics. Finally, investigators must pre-specify a limit to the number of experimental agents that may be evaluated simultaneously and ultimately added to the platform in order adjust the thresholds to control for multiplicity.

6.3 Threshold calibration

Simulation was applied to calibrate the design parameters for a screening trial using the platform design. Decision thresholds, θ and ϕ, were calibrated to maximize power for detecting at least one effective experimental therapy while controlling familywise type I error rate (FWER) at 0.10.

Table 1. Simulated frequentist familywise type I error rate and power for a platform trial designed to test a maximum of five experimental arms resulting from 20 combinations of the decision thresholds.

	$\theta\psi$				
$\phi\psi$	0.62	0.64	0.66	0.68	0.70
Familywise Type I error rate					
0.001	0.133	0.131	0.099	0.092	0.089
0.005	0.137	0.127	0.095	0.087	0.085
0.01	0.122	0.116	0.090	0.085	0.084
0.05	0.111	0.080	0.075	0.074	0.072
Power					
0.001	0.841	0.822	0.809	0.795	0.765
0.005	0.830	0.805	0.795	0.786	0.762
0.01	0.823	0.784	0.781	0.779	0.758
0.05	0.767	0.720	0.711	0.699	0.680
Expected sample size under the null scenario 0					
0.001	316.8	307.1	303.7	302.2	298.8
0.005	282.0	273.0	267.1	264.3	263.7
0.01	258.7	249.9	245.9	242.4	241.5
0.05	190.2	184.7	181.4	176.3	176.2
Expected sample size under the alternative scenario 1					
0.001	338.9	335.3	330.0	328.2	327.8
0.005	312.1	306.9	303.4	300.7	297.2
0.01	297.8	292.4	284.6	281.1	281.0
0.05	240.9	232.9	228.4	227.0	224.8

Shaded regions delineate threshold combinations that yield a design that controls the familywise type I error rate at ≤ 0.10.

Delayed entry into the platform induces comparisons with increased sample size for control, thereby providing additional information for estimating the control response rate, $\pi_0|r_0$. Thus, threshold calibration need only consider the case wherein all five experimental arms are available at baseline. Table 1 provides FWER and power as well as the expected sample size that was obtained under the null and alternative scenario 1 for a subset of 20 combinations of $\theta\psi$ and ϕ.

The FWER reflects the probability that the design concludes superiority (over control) for any of the five experimental arms under the global null scenario wherein the true response rate is identically 0.2 for control and all five experimental arms. For fixed $\phi\psi(\theta)$, increasing $\theta\psi(\phi)$ decreases FWER. The table's shaded regions delineate threshold combinations that yield a platform design that controls FWER < 0.10. The power reflects the probability of detecting a two-fold increase in the response rate for exactly one experimental therapy (a response rate of 0.4) when the other four experimental therapies are truly equivalent to control. For fixed $\phi\psi(\theta)$, decreasing $\theta\psi(\phi)$ increases the design's power. Among the admissible choices, the threshold combination $\phi = 0.001$ and $\theta = 0.66$ is most powerful. One should note, however, that moderate reductions in sample size may be effectuated through relatively modest FWER inflation. For example, the threshold combinations (ϕ, θ) of (0.005, 0.64) and (0.01, 0.62) reduce the mean sample size under the null scenario by 10% and 15%, respectively, while controlling FWER < 0.13 and maintaining power > 0.80. Decision thresholds for the two-arm sequential PP design were similarly

Table 2. Trial operating characteristics obtained when screening agents one-at-a-time using a sequence of five two-arm sequential predictive probability designs.

Scenario		True response rate	Average no. patients		Probability		Average	
			assigned	respond	not dropped for futility	all null arms dropped for futility	total sample size	total duration (in years)
0	control 1-5	0.2	49.8	10.0	–	0.901	498	4.15
	exp. 1-5	0.2		10.0	0.021			
1	control 1-4	0.2	49.8	10.0	–	0.918	538	4.46
	exp. 1-4	0.2		10.0	0.021			
	control 5	0.2	69.5	13.9	–			
	exp. 5	0.4		27.8	0.809			
2	control 1-3	0.2	49.8	10.0	–	0.938	579	4.80
	exp. 1-3	0.2		10.0	0.021			
	control 4-5	0.2	69.5	13.9	–			
	exp. 4-5	0.4		27.8	0.809			
3	control 1	0.2	32.7	6.5	–	0.980	572	4.77
	exp. 1	0.1		3.3	0.00			
	control 2	0.2	49.8	10.0	–			
	exp. 2	0.2		10.0	0.023			
	control 3	0.2	63.8	12.8	–			
	exp. 3	0.3		19.1	0.334			
	control 4	0.2	69.4	13.9	–			
	exp. 4	0.4		27.7	0.817			
	control 5	0.2	70.0	14.0	–			
	exp. 5	0.5		35.0	0.99			

calibrated to attain power ≥ 0.80 while controlling FWER ≤ 0.10 over the course of five successive trials.

6.4 Operating characteristics

Table 2 provides the resulting operating characteristics for the two-arm sequential PP design. Scenario 0 represents the global null case. Because trials with study arms that have identical true response probabilities are exchangeable, the numerical values of operating characteristics that are reported for a set of experimental arms characterize the identical value that is attained for each individual arm. For example, under Scenario 0 each comparison to the five experimental arms attains type I error of 0.021. Here we see that an average null trial enrolled approximately 49.8 patients per arm, requiring 498 patients to screen five agents with FWER controlled at 0.1. The total sample size is increased by 40 in the presence of four null agents and one agent that is truly superior (as depicted in scenario 1). This resulted from the fact that one of the trials (the fifth trial in our simulation) tends to elude early stopping, enrolling an additional 20 patients per arm on average. In the presence of two identically effective agents (presented in scenario 2), the sequential trials required approximately 579 patients on average. Scenario 3 considers performance using experimental

Table 3. Trial operating characteristics resulting from the proposed platform design when all five experimental therapies are available at baseline.

Scenario		True response rate	Average no. patients		Probability		Average	
			assigned	respond	not dropped for futility	all null arms dropped for futility	total sample size	total duration (in years)
0	control	0.2	62.4	12.5	–	0.906	304	2.53
	exp. 1-5	0.2	48.3	9.7	0.026			
1	control	0.2	69.3	13.9	–	0.920	330	2.75
	exp. 1-4	0.2	47.9	9.6	0.024			
	exp. 5	0.4	69.0	27.6	0.809			
2	control	0.2	69.9	14.0	–	0.944	350	2.92
	exp. 1-3	0.2	47.5	9.5	0.022			
	exp. 4-5	0.4	69.1	27.6	0.802			
3	control	0.2	70.0	14.0	–	0.972	350	2.91
	exp. 1	0.1	30.3	3.0	0.00			
	exp. 2	0.2	47.5	9.5	0.028			
	exp. 3	0.3	62.9	18.9	0.312			
	exp. 4	0.4	69.2	27.7	0.815			
	exp. 5	0.5	69.9	35.0	0.982			

response rates that vary from 0.1 to 0.5. Trials 1–5 enrolled 32.7, 49.8, 63.8, 69.4, and 70.0 patients on average, respectively. Increasing the response rate for an experimental arm reduces the chance of futility stopping, thereby yielding a larger trial. For all scenarios, the probability that an experimental arm with $\pi \geq 0.4$ eludes early stopping for futility exceeds 0.8. Moreover, the design provides strong control of FWER at ≤ 0.1 (as evident in the sixth column of Table 2).

Tables 3 and 4 summarize the operating characteristics that result from implementation of the proposed screening platform design. Table 3 considers the case wherein all five experimental agents are available at baseline. Table 4 presents results in the presence of delayed study entry, such that a new experimental arm is added to the platform after the previously enrolled agent has accumulated 10 patients. Because the standard of care therapy is maintained as a treatment option for the duration of the entire study period, in the presence of delayed entry (Table 4) the number of patients assigned to the control arm may exceed $N_{\max}$, thereby increasing the total sample size for the platform when compared to Table 3. Additionally, Table 5 compares the aggregate mean response rate for the trial and the proportion of patients assigned to arms for which $\pi > \pi_0 + \delta$.

When compared to conventional sequential two-arm trials with commensurate frequentist size, power, and FWER; trial consolidation using the proposed platform design improved the overall response rate, while requiring fewer patients and a shorter duration. Specifically, the platform designs required 39% fewer patients (304 versus 498) to screen five null agents in Scenario 0 when compared to the one-at-a-time approach when all five agents were available at baseline. The platform resulted in a 36% reduction with delayed entry (w.d.e.). For scenarios 1, 2, and 3, the platform approach yielded reductions of 39% (36% w.d.e.), 40% (36% w.d.e.), and 39% (32% w.d.e.), respectively, with increased gains in efficiency when all agents are available at baseline.

Table 4. Trial operating characteristics resulting from the proposed platform design in the presence of delayed study entry.

Scenario		True response rate	Average no. patients		Probability		Average	
			assigned	respond	not dropped for futility	all null arms dropped for futility	total sample size	total duration (in years)
0	control	0.2	90.4	18.1	–	0.917	318.6	2.65
	exp. 1-5	0.2	45.6	9.1	0.019			
1	control	0.2	91.6	18.3	–	0.935	342.1	2.85
	exp. 1,3,4,5	0.2	45.4	9.1	0.019			
	exp. 2	0.4	69.0	27.6	0.797			
2	control	0.2	95.3	19.1	–	0.952	370.2	3.09
	exp. 1,4,5	0.2	45.7	9.1	0.021			
	exp. 2-3	0.4	68.9	27.6	0.805			
3	control	0.2	109.9	22.0	–	0.983	387.7	3.23
	exp. 1	0.1	30.6	3.1	0.00			
	exp. 2	0.2	46.3	9.3	0.025			
	exp. 3	0.3	62.0	18.6	0.316			
	exp. 4	0.4	68.9	27.6	0.828			
	exp. 5	0.5	69.9	35.0	0.987			

Table 5. Operating characteristics that result from screening five experiment agents using sequential randomized two-arm trials, platform design, and platform design with delayed entry.

Design property	Design	Scenario			
		0	1	2	3
Proportion of patients assigned to arm with success rate $\pi \geq 0.3$	Sequential two-arm trials	0	0.13	0.24	0.36
	Platform	0	0.21	0.39	0.58
	Platform with delayed entry	0	0.20	0.37	0.52
Mean trial response rate	Sequential two-arm trials	0.20	0.23	0.25	0.27
	Platform	0.20	0.24	0.28	0.31
	Platform with delayed entry	0.20	0.24	0.27	0.30

Table 5 shows that the proportions of patients assigned to effective arms ($\pi \geq 0.3$) were 36%, 58%, and 52% for the three designs in Scenario 3. The corresponding mean overall response rates were 0.27, 0.31, and 0.30, respectively. Therefore, the platform design increased the overall response rate by 15% (11% w.d.e.), on average.

7 Trial conduct

We recognize that the computation of Bayesian posterior distributions and PP can be time consuming which may impede trial conduct. However, after having specified the three design arguments, futility

monitoring boundaries for the PP-based platform design presented in this article can be computed prior to initiating the study. This facilitates continuous screening of competing agents without the need to implement Bayesian computation at interim analyses during the trial. The supporting web materials provide example futility monitoring tables that result from implementation of the screening platform used in the simulation study. The elements of Supporting Tables provide the minimum number of therapeutic responses that are required in order for an experimental arm to avoid early stopping under the PP-based futility decision rule. Supporting Table S1 provides monitoring boundaries without delayed entry. Thus, the sample size per arm (represented by column) is assumed to be identical for experimental and control arms. Here for example, we see that after observing exactly 11 patients on each arm, at least one responder is required in order to continue enrollment to the experimental arm given that four patients have responded to the control therapy. Supporting Table S2 depicts the futility boundaries in the presence of sample size imbalance, after 35 patients have been treated with control and up to 12 patients have been treated with the experimental therapy. Given less than 10 treatment successes for control, the jth experimental agent will continue to enroll an additional patient. However, given that 14 patients have responded to the control therapy, at least one response is required if $8 \leq n_j \leq 11$. Two responders are required after observing responses from 12 patients treated with the experimental therapy. If the control therapy has induced responses in all 35 patients, the platform should stop due to the fact that it would be highly unlikely to demonstrate an improvement in the response rate of size $\delta = 0.1$ for any therapy. Upon specification of the design parameters, such tables can be generated to facilitate conduct of the clinical trial.

8 Discussion

The process of screening agents one-at-a-time under the current clinical trials system suffers from several deficiencies that could be addressed in order to extend financial and patient resources. Efficiency is especially important oncology settings, wherein trials already suffer from low enrollment and an unacceptably high rate of failure in phase III. Randomized screening platforms consolidate resources and improve the quality of the primary drug development decision regarding whether to proceed to phase III following phase II. When compared to sequentially conducted, randomized two-arm trials, screening platforms have the potential to yield considerable reductions in cost by requiring fewer patients and better avoiding expensive confirmatory phase III studies for ineffective agents. We demonstrated that these gains are attainable without sacrificing the statistical properties of the treatment comparisons. Because the design is amenable to evaluation of stopping boundaries using look-up tables, the need to use advanced statistical software to conduct the trial can be avoided. A platform-based approach to phase II drug development would be most useful in environments where multiple emerging therapies would be compared to a common established standard of care therapy.

Statistical methods for monitoring futility in the context of platform design derive most naturally from PP using the Bayesian paradigm, but not limited to the tools presented here. The reader should note that our simulation studies compared trial durations for platform designs to two-arm trials that were conducted in sequence. Therefore, the reported reductions in time for the platform approach could be lessened when compared to those obtained for a collection of two-arm trials that are conducted simultaneously. However, the platform approach would maintain the reported reductions in total sample size for this comparison. Our proposed platform design with PP monitoring could be viewed as a snapshot of a "perpetual trial" [p.61].[36] In fact, one would only need to establish criteria for periodically replacing the control arm to effectuate one type of perpetual trial using the tools presented here.

Our simulation study reported operating characteristics for the platform design that were obtained using $\phi = 0.001$ and $\theta = 0.66$. While this threshold pair provided the optimal design on the basis of frequentist power and size alone, we can see from Table 1 that changes in ϕ will impact the design's expected sample size, with smaller values requiring a larger trial to attain power of 0.8. Thus, effective agents are more likely to be dropped for futility under the continuous monitoring scheme with larger values of ϕ. In practice one might endeavor to choose these thresholds by considering frequentist power and size conjointly with the design's expected sample size.

There are several practical issues that one needs to consider before implementing a platform design. The effectiveness of any sequential trial monitoring scheme is impacted by the rate of patient accrual and the duration of time that is required before patient outcomes can be ascertained. More rapid recruitment and/or prolonged delay between the time of enrollment and assessment of treatment response diminish the extent to which one can limit enrollment to ineffective experimental agents. When the control arm is retained for the duration of the platform, bias due to population-drift could affect treatment comparisons, in particularly at latter stages of the trial. One can induce robustness to population-drift by limiting the extent to which each control patient can influence each interim analysis on the basis of their time of enrollment. For example, interim analyses at trial time t could include only those control patients who were enrolled within some pre-determined interval of time (perhaps within 1–2 years) from t. A more stringent approach would restrict comparisons with experimental agent j to only those patients who were randomized to the control therapy after agent j was incorporated into the platform. We illustrate this modified design schema of such a platform trial in Supporting Figure S1. Both approaches represent specific cases of a general beta-binomial model that utilizes a power prior for π that determines the extent to which subjects that receive the same therapy are considered "exchangeable" using some mapping of enrollment time to the unit interval. We are now investigating platform designs that utilize this power prior framework adaptively by assessing the enrollment trends of important clinical/prognostic covariates and specifying the power parameter in relation to the extent of "evidence" for population-drift.

While we evaluated platform designs that effectuate strong control of FWER at ≤ 0.10, control of familywise type I error is often neglected in practice when multiple comparative hypothesis tests arise in the context of a multi-arm study. In fact, the authors of a recent systematic review of multi-arm trials published in four major medical journals during 2012,[37] which found that correction for multiple comparisons was absent in 45% (9/20) of exploratory and 54% (21/39) of confirmatory trials, explicitly advocate for guidelines from regulatory bodies to establish circumstances when multiplicity adjustment should be considered requisite for trials that evaluate multiple therapies.

The general concepts conveyed in this article could be applied to devise platform trials that utilize other types of endpoints. Due to the luxury of analytical tractability, adaptations to settings that involve Gaussian likelihoods are relatively straightforward. Platform designs could also be based on time-to-event endpoints using tractable exponential-gamma models as well as more flexible semi-parametric hazard or accelerated failure time models. Depending on the event rate versus recruitment rate trade-off, platform designs with continuous monitoring might be limited in settings wherein time to treatment failure is used as the basis for treatment comparison, but rather better suited to compare surrogate endpoints of clinical efficacy that are observable shortly following treatment.

Recent efforts to enhance the drug development process have largely focused on the development of seamless phase II/III trials, which integrate phase II and III into a single trial without stopping patient accrual. For example, Inoue et al.[38] found that seamless phase II/III designs may yield

reductions in average sample size ranging from 30% to 50% when compared to more conventional designs with identical frequentist properties. More recently, several types of phase II/III designs have been proposed. Kimani et al.[39] proposed a dose-selection procedure in the context of an adaptive phase II/III trial with binary endpoints that incorporates the dose-response relationships into the treatment comparison. Stallard[40] considered strong control of the familywise type I error rate when short-term endpoints are used for the treatment selection at the phase II stage. Given the extent to which platform design promises to enhance the efficiency of the drug screening process, we believe that the phase II/III design paradigm could be further enhanced using the methods described here. Implementation would require increased multi-institutional collaboration with industry, however. The Lung Cancer Master Protocol (Lung-MAP) study,[41] a multi-institutional, multi-cooperative group phase II/III trial devised to individualize treatment for patients with squamous cell lung cancer, demonstrates that such partnerships are necessary and feasible.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Brian P. Hobbs and J. Jack Lee were supported in part by the University of Texas M.D. Anderson's Cancer Center Support Grant NIH P30 CA016672. Nan Chen was fully funded by MD Anderson internal funds.

References

1. Nass SJ, Moses HL and Mendelsohn J. Institute of Medicine: a national cancer clinical trials system for the 21st century: reinvigorating the NCI Cooperative Group Program. Washington, DC: National Academies Press, 2010.
2. Mayfield E. Combining targeted cancer therapies: much promise, many hurdles. *NCI Cancer Bulletin* 2011; **8**(10): 6.
3. Sutter S and Lamotta L. Cancer drugs have worst phase III track record. *Internal Medicine News Digital Network* 2011.
4. DiMasi JA, Reichert JM, Feldman L, et al. Clinical approval success rates for investigational cancer drugs. *Clinical Pharmacology and Therapeutics* 2013; **94**: 329–335.
5. Pharmaceutical Research and Manufacturers of America, *771 medicines in development for cancer*. http://www.phrma.org/research/cancer
6. Pharmaceutical Research and Manufacturers of America. *Drug discovery and development*, http://www.phrma.org/sites/default/files/pdf/rdbrochure022307.pdf
7. Seymour L, Ivy SP, Sargent D, et al. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the National Cancer Institute investigational drug steering committee. *Clinical Cancer Research* 2010; **16**: 1764–1769.
8. Lee JJ and Chu CT. Novel statistical models for NSCLC clinical trials. In: Roth JA, Hong WK and Komaki RU (eds) *Lung cancer*. Hoboken, New Jersey: Wiley-Blackwell, 2014, pp.488–503.
9. Freidlin B, Korn E, Gray R, et al. Multi-arm clinical trials of new agents: some design considerations. *Clinical Cancer Research* 2008; **14**: 4368–4371.
10. Barthel FMS, Parmar MKB and Royston P. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design—a reanalysis of 4 trials. *Trials* 2009; **10** article number 21).
11. Parmar MKB, Carpenter J and Sydes MR. More multiarm randomised trials of superiority are needed. *The Lancet* 2014; **384**: 283–284.
12. Wason JMS and Jaki T. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* 2012; **31**: 4269–4279.
13. Wason JMS and Trippa L. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine* 2014; **33**: 2206–2221.
14. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* 1961; **13**: 346–353.
15. Simon R. Optimal 2-stage designs for phase-II clinical trials. *Controlled Clinical Trials* 1989; **10**: 1–10.
16. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**: 143–151.
17. Chang MN, Therneau TM, Wieand HS, et al. Designs for group sequential phase II clinical trials. *Biometrics* 1987; **43**: 865–874.
18. Yao TJ, Begg CB and Livingston PO. Optimal sample size for a series of pilot trials of new agents. *Biometrics* 1996; **52**: 992–1001.
19. Thall PF and Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50**: 337–349.
20. Lee JJ and Liu DD. A predictive probability design for phase II cancer clinical trials. *Clinical Trials* 2008; **5**: 93–106.
21. Mandrekar SJ and Sargent DJ. Randomized phase II trials: time for a new era in clinical trial design. *Journal of Thoracic Oncology* 2010; **5**: 932–934.

22. Simon R, Wittes RE and Ellenberg SS. Randomized phase II clinical trials. *Cancer Treatment Reports* 1985; **69**: 1375–1381.
23. Whitehead J. Designing phase ii studies in the context of a programme of clinical research. *Biometrics* 1985; **41**: 373–383.
24. Lee JJ and Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. *Journal of Clinical Oncology* 2005; **23**: 4450–4457.
25. Yin G, Chen N and Lee JJ. Phase II trial design with Bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society, Series C* 2012; **61**: 219–235.
26. Whitehead J. Sample sizes for phase ii and phase iii clinical trials: an integrated approach. *Statistics in Medicine* 1986; **5**: 459–464.
27. Stallard N. Optimal sample sizes for phase ii clinical trials and pilot studies. *Statistics in Medicine* 2012; **31**: 1031–1042.
28. Wason JMS, Jaki T and Stallard N. Planning multi-arm screening studies within the context of a drug development program. *Statistics in Medicine* 2013; **32**: 3424–3435.
29. Simon R. Advanced clinical trial educational session. In: *The American Society of Clinical Oncology Meetings*, Chicago, 4 June 2010.
30. Liu P, LeBlanc M and Desai M. False positive rates of randomized phase ii designs. *Controlled Clinical Trials* 1999; **20**: 343–352.
31. Therasse P, Arbuck S, Eisenhauer E, et al. New guidelines to evaluate the response to treatment in solid tumors: European organization for research and treatment of cancer, national cancer institute of the united states, national cancer institute of Canada. *Journal of the National Cancer Institute* 2000; **92**: 205–216.
32. Lara P, Redman M, Kelly K, et al. Disease control rate at 8 wk predicts clinical benefit in advanced non-small-cell lung cancer: results from Southwest Oncology Group randomized trials. *Journal of Clinical Oncology* 2009; **26**: 463–467.
33. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery* 2011; **1**: 44–53.
34. Prowell T and Pazdur R. Pathological complete response and accelerated drug approval in early breast cancer. *New England Journal of Medicine* 2012; **366**: 2438–2441.
35. Wathen JK and Thall PF. Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statistics in Medicine* 2008; **27**: 5586–5604.
36. Wizemann T, Altevogt BM and Claiborne AB. *Institute of Medicine forum on drug discovery, development, and translation; Institute of Medicine forum on medical and public health preparedness for catastrophic events. Advancing regulatory science for medical countermeasure development: Workshop summary*. Washington, DC: National Academies Press, 2011.
37. Wason JMS, Stecher L and Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* 2014; **15**: 364.
38. Inoue LYT, Thall P and Berry DA. Seamlessly expanding a randomized phase ii trial to phase iii. *Biometrics* 2002; **58**: 823–831.
39. Kimani PK, Stallard N and Hutton JL. Dose selection in seamless phase ii/iii clinical trials based on efficacy and safety. *Statistics in Medicine* 2009; **28**: 917–936.
40. Stallard N. A confirmatory seamless phase ii/iii clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; **29**: 959–971.
41. Steuer C, Papadimitrakopoulou V, Herbst R, et al. Innovative clinical trials: the lung-map study. *Clinical Pharmacology & Therapeutics* 2015; **97**(5): 488–491.