CLINICAL INVESTIGATION

Using simulation to optimize adaptive trial designs: applications in learning and confirmatory phase trials

Failures in Phase III clinical trials have a major impact on escalating drug development costs. Clinical trial simulation can improve decisions made in learning Phase I and Phase II studies, leading to better designs and reduced chance of failure in high-cost, confirmatory Phase III trials. In adaptive design trials – which use early findings to modify the trial toward the most promising dose, sample size or patient population – simulation is instrumental in guiding selection of optimal design modifications. In this article, the authors present case studies demonstrating the benefits of simulations to identify maximum tolerated dose in Phase I, define dose–response relationships in Phase II, and determine the best type and timing of interim analyses in Phase III.

Keywords: adaptive trial design • clinical trial simulation • continual reassessment method • group sequential design • sample size re-estimation

Jürgen Hummel*,[1],
Song Wang[2]
& John Kirkpatrick[1]
[1]Biostatistics, PPD, Fleming House, Phoenix Crescent, Strathclyde Business Park, North Lanarkshire, ML4 3NJ, Scotland, UK
[2]Biostatistics, PPD, 7551 Metro Center Drive, Suite 300, Austin, TX 78744, USA
*Author for correspondence:
Tel.: +44 1698 575228
jurgen.hummel@ppdi.com

High failure rates in clinical trials are a major contributor to the escalating costs of drug development, now estimated to be as high as $1.8 billion per approved drug [1]. According to a 2014 survey of clinical development success rates, 36% of experimental agents entering Phase I fail to advance to Phase II, 68% of Phase II entries fail to advance to Phase III and 17% of Phase III entries fail to earn regulatory approval [2].

Failures in large confirmatory Phase III trials have the greatest impact on research investment. In oncology, for example, only 34% of Phase III trials achieve statistical significance in their primary end points [3]. A 2013 analysis suggests that for big pharmaceutical companies with development pipelines large enough to achieve 8–13 drug approvals annually, the cost per approval is closer to $5 billion due to financial losses resulting from numerous failures [4]. Acknowledging that the current development model is not sustainable, sponsors and regulators are pursuing new strategies and methodologies to make clinical trials more informative and efficient.

One relatively new approach is the mathematical modeling and simulation of clinical trials, which can be used to simulate study results under a variety of scenarios and guide the selection of study design parameters – such as randomization ratio, sample size, duration and number of dosing levels – that are most likely to result in a trial that answers the key study question. Clinical trial simulation can be used for almost any type of study design, but the technique is particularly well suited for adaptive designs where the decision-making processes are often complex and the design-operating characteristics are difficult to evaluate using conventional methods. By quantifying potential outcomes of various design options, simulations enable sponsors to make more accurate, evidence-based decisions at each step of the drug evaluation process, culminating in a greater likelihood of success in Phase III.

In this article, the authors discuss the use of simulation studies to optimize adaptive trial designs. Case studies are presented to illustrate simulations in Phase I to better identify the maximum tolerated dose (MTD), simulations in Phase II aimed at better defining dose–response relationships and simulations in Phase III to determine the right sample

10.4155/CLI.15.14 © 2015 Future Science Ltd

ISSN 2041-6792

size by allowing early stopping for success and futility and building in sample size re-estimation.

Modeling & simulation in clinical trial design

The mathematical modeling of biological phenomena, including chemical reactions, drug pharmacokinetics and nerve conduction, is not a new concept. However, until the late 1980s, these models were generally limited to relatively straightforward, deterministic equations that could be solved by hand. With the advent and exponential proliferation of available computing power, biological models have become increasingly complex and now include stochastic (random) elements that allow for individual variability in response (Figure 1).

Stochastic models are the cornerstone of clinical trial simulation. As a simplified (and fictitious) example, consider a Phase II dose-ranging study for a new antihypertensive agent. Based on the results from preclinical and Phase I studies, researchers have estimated the drugs most likely dose–response curve (Figure 2). Assuming that toxicity is not dose-dependent in this range, it seems clear from Figure 2 that the most effective dose is likely to be 50 mg. But what is the probability that a clinical trial, whose results will vary by individual, will select the 'correct' answer with a limited sample size?

The purpose of clinical trial simulation is to answer this question before the trial begins. Consider a simplified example in which we assume that the proposed antihypertensive trial uses a fixed design that randomizes 80 patients to each dose illustrated in Figure 2 and to the control (total n = 400). For each modeled patient receiving a given dose, the simulation will randomly select a response within the predetermined range (based on the SE bars in Figure 2). After the data from all 400 patients have been simulated, the software analyzes the results using the same statistical methods and criteria that will be used to determine the minimum effective dose in the 'real' clinical trial. This simulation process is repeated (usually between 1000 and 100,000 times), and the output is the percentage of simulations that select a specific dose as the minimum effective dose (Figure 3). If success is defined as determining the minimum effective dose of 50 mg, the example trial has a <50% chance of achieving its goal – poor odds for generating reliable information on which to base dose selection for Phase III development.

The true power of clinical trial simulation is the ability of this approach to compare the likely outcomes of a number of scenarios and help guide the selection of the best trial design based on what is already known, or assumed to be known, about the drug. For example, preclinical data may inform Phase I; Phase I data may inform Phase II; Phase II data may inform Phase III. Decisions that traditionally are made based on research

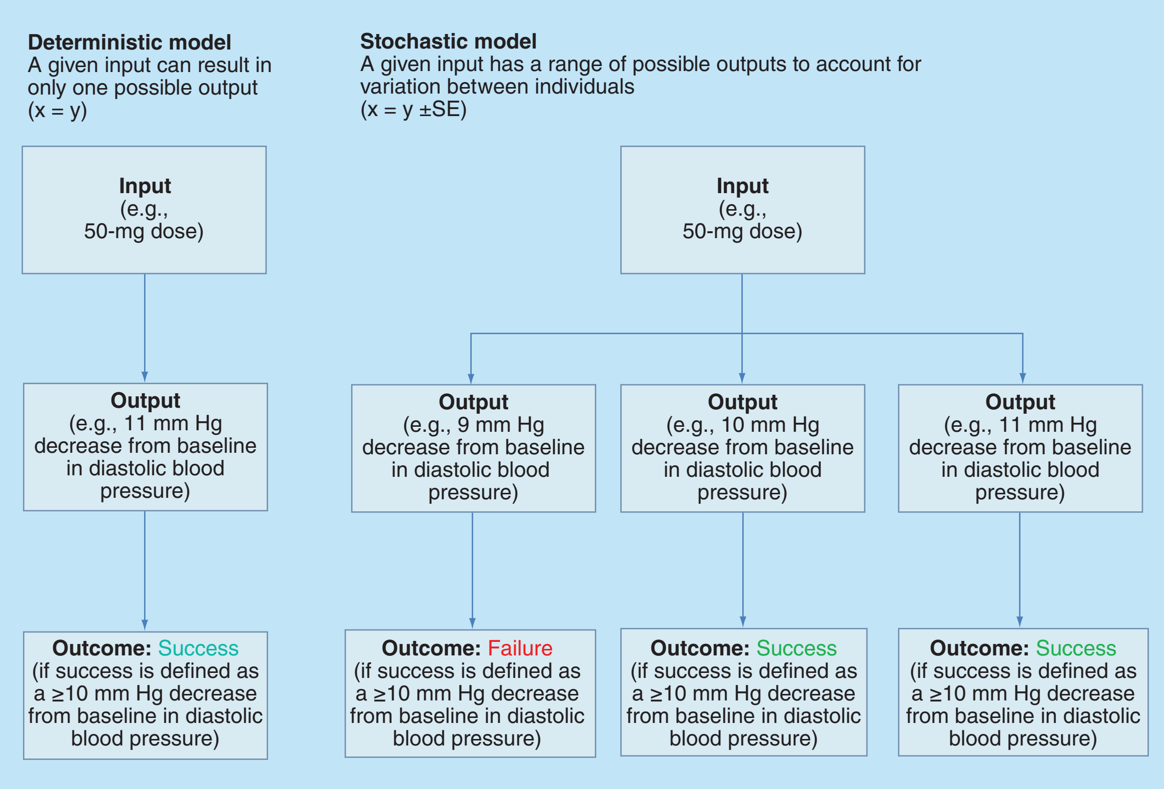
Deterministic model
A given input can result in only one possible output
(x = y)
Stochastic model
A given input has a range of possible outputs to account for variation between individuals
(x = y ±SE)
Input (e.g., 50-mg dose)
Input (e.g., 50-mg dose)
Output (e.g., 11 mm Hg decrease from baseline in diastolic blood pressure)
Output (e.g., 9 mm Hg decrease from baseline in diastolic blood pressure)
Output (e.g., 10 mm Hg decrease from baseline in diastolic blood pressure)
Output (e.g., 11 mm Hg decrease from baseline in diastolic blood pressure)
Outcome: Success (if success is defined as a ≥10 mm Hg decrease from baseline in diastolic blood pressure)
Outcome: Failure (if success is defined as a ≥10 mm Hg decrease from baseline in diastolic blood pressure)
Outcome: Success (if success is defined as a ≥10 mm Hg decrease from baseline in diastolic blood pressure)
Outcome: Success (if success is defined as a ≥10 mm Hg decrease from baseline in diastolic blood pressure)

Figure 1. Example of a deterministic versus stochastic model.
SE: Standard error.

experience or guesswork can be considered based on quantification of a parameter's impact on the number of patients required, study duration and cost. Incremental improvement at each step of evaluation improves the design – and probability of success – of the next step.

The typical goal of clinical trial simulation is to identify a design that has a high probability of success based on the most likely conditions (i.e., the expected 'U'-shaped dose–response curve in Figure 2) but which can also perform well, or at least acceptably, under more extreme conditions (i.e., a linear or sigmoidal dose–response curve) if these conditions happen to occur.

The US FDA has endorsed clinical trial simulation, stating that the results 'could reduce the risk and cost of human testing by helping product sponsors make more informed decisions on how to proceed with product testing and when to remove a product from further development' [5]. While the European Medicines Agency (EMA) has no official publications on clinical trial simulation as yet, the EMA is receptive to the use of simulations.

As in any well-designed scientific experiment, the aims of a simulation study and its methods should be defined in advance [6]. However, a simulation study often highlights either areas worthy of further investigation or behaviors of the simulated design that are unexpected. Therefore, it is important that the simulation plan allows for iterations, with the results of each stage influencing the aims and methods of the next. The narratives in this paper focus on this iterative process.

Simulation in adaptive trial designs: optimizing modifications

In its 2010 draft guidance, FDA defined an adaptive design clinical study as 'a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of the data (usually interim data)' [7]. FDA-cited modifications are shown in Box 1. In adaptive approaches, early findings can be used to redirect the trial toward the most promising doses, disease indications or patient populations. For example, if an interim analysis finds that the lowest dose is ineffective, evaluation of that group can be halted and the patients reallocated to higher, possibly more effective doses.

When choosing an adaptive study design, the overall goal is to increase the likelihood that the clinical trial will succeed in correctly answering the question it was intended to address. Even a study that is stopped early for futility can be considered successful in the sense that it has answered the key research question while minimizing resource use. Adaptive studies may, in some situations, accelerate study time lines and reduce costs. However, the complexity of these trials also frequently lengthens protocol development cycles, increases the need for statistical expertise and extends regulatory review times [8].

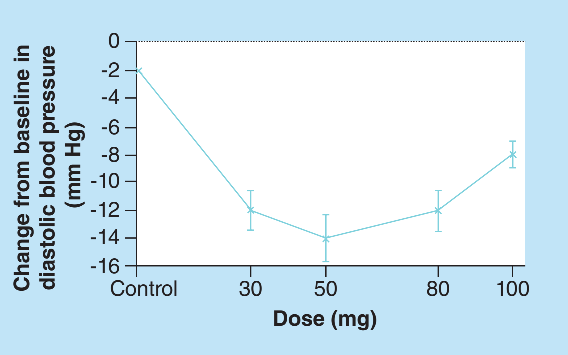

Figure 2. Fictitious example of a dose–response curve ±standard error for a new antihypertensive drug.

The intrinsic complexity of adaptive trial designs makes clinical trial simulation an effective and efficient tool to choose the adaptive elements that will optimize the overall design and improve the probability of success. Box 2 shows examples of design choices, simulation plan decisions and operating characteristics (outputs) commonly used in simulations for adaptive trials.

Based on the significant role that clinical trial simulation plays in the design of adaptive trials, regulatory agencies suggest that sponsors include simulation results when seeking feedback on adaptive studies [7]. Because there is rarely a single 'best' design for an adaptive trial, written explanations of the rationale behind the sponsor's choices can also be valuable during the review process.

The following case studies illustrate the power of simulation to inform clinical trial design and increase probabilities of success in Phase I, Phase II and Phase III. Two simulation software programs were used: Berry Consultants' Fixed and Adaptive Clinical Trial Simulator (FACTS™, Austin, TX)

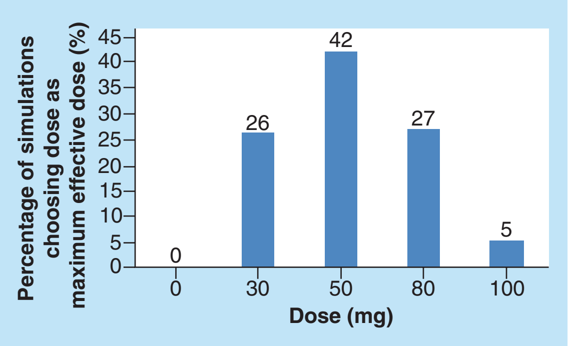

Figure 3. Percentage of simulations choosing a specific dose as the minimum effective dose based on the fictitious dose response curve in Figure 2 and a fixed study design with 80 patients randomized to each dose and to control.

Box 1. US FDA-specified examples of possible modifications during an adaptive trial†.

- Study eligibility criteria (either for subsequent study enrollment or for a subset selection of an analytic population)
- Randomization procedure
- Treatment regimens of the different study groups (e.g., dose level, schedule, duration)
- Total sample size of the study (including early termination)
- Concomitant treatments used
- Planned schedule of patient evaluations for data collection (e.g., number of intermediate time points, timing of last patient observation, duration of patient study participation)
- Primary end point (e.g., which of several types of outcome assessments, which time point of assessment, use of a single vs composite end point or the components included in a composite end point)
- Selection and/or order of secondary end points
- Analytic methods to evaluate the end points (e.g., covariates of final analysis, statistical methodology, type I error control)

†Data taken from [6].

for Phase I and II simulations, and Cytel Corporation's (Cambridge, MA) EAST® for Phase III simulations. Customized programming using software such as R or SAS can provide additional flexibility and therefore be very powerful for simulations, and can therefore be very powerful tools for examining nonstandard situations. However, specialized software packages are easier to use in standard situations and

Box 2. Examples of design options, simulation plan decisions, and operating characteristics commonly used in clinical trial simulation for adaptive trials.

Design options

- Treatments
  - Duration
  - Number of dosing levels
  - Choice of control(s)
- Randomization
  - Fixed equal
  - Fixed unequal
  - Dynamic allocation (e.g., response adaptive)
  - Minimization (e.g., biased coin)
- Sample size/power (design can fix one or the other)
  - Study objective(s)
  - Primary end point
  - Definition of clinical benefit and effect size
- Interim analysis
  - Purpose (e.g., early stopping criteria for efficacy or futility, sample size recalculation, discontinuation of treatment arms)
  - Number and timing

Simulation plan decisions

- Design components used to guide the simulation
  - Choice of statistical modeling and analysis methods
  - Definition of success (e.g., correct choice of dose, statistically significant difference between active and control groups)
  - Decision rules at interim analyses (e.g., stopping boundaries for futility/ success/arm dropping, dose escalation rules)
- Scenarios used to evaluate design performance
  - Range and shape of possible dose–response or dose–toxicity curves (e.g., linear, sigmoidal, Emax)
  - Variability in dose–response or dose–toxicity curves

Operating characteristics (outputs)

  - Probability of success (based on definition of success)
  - Sample size (average, minimum, maximum)
  - Study duration (average, minimum, maximum)
  - Estimated power and/or type I error rate (particularly important for Phase III trials)

provide useful insight. FACTS is an industry-leading platform particularly powerful when simulating learning-phase trials. EAST has special utility when the goal of simulations is to compare various designs for confirmatory studies.

Phase I case study: simulations to determine MTD

A central goal of Phase I evaluation is to determine the MTD of a drug entering clinical development. Failure to choose the correct MTD in Phase I means that the wrong dose (or doses) will be studied in Phase II and Phase III trials.

In this Phase I study of an oncology drug, the sponsor planned to use a standard 3+3 design to determine MTD. In the 3+3 approach, patients are treated in cohorts of three and the results from each cohort are used to decide whether or not dose escalation will continue [9].

The 3+3 approach is not particularly efficient [10]. Simulations were suggested to find a more optimal study design based on the Neuenschwander continual reassessment method (NCRM) [11]. The NCRM approach updates the dose–toxicity curve after each patient completes the study. One patient is enrolled sequentially at each dose level until a dose-limiting toxicity is observed. At that time, a three-patient cohort is initiated on the MTD predicted by the model using the data collected to date. The trial is stopped when a prespecified number of cohorts have been treated at the MTD, or when the planned maximum sample size has been attained. Other approaches, such as the modified toxicity probability interval design [12], are also available; for comparison, see discussions presented by Ji and Pan [13].

For this Phase I oncology trial, the dose range to be tested was 0.1–2.5 mg/kg (based on manufacturing capabilities), the maximum sample size was 36 and the target toxicity range at the MTD was 20–33%. In addition to the 3+3 design, simulations were conducted for three NCRM designs that differed in the number of confirming cohorts treated at the MTD, and in their tolerance of overdosing risk (Table 1).

Based on preclinical data, the experimental drug was thought to be relatively nontoxic. The MTD was unlikely to be in the lower end of dose range, and it was possible that it was higher than the maximum feasible dose tested.

To cover a range of possible outcomes, simulations were performed using FACTS for six potential dose–toxicity curves (Figure 4). The high-toxicity profiles (H1 and H2) are the least likely to occur, but it was important to evaluate how each trial design performed under both probable and extreme conditions. The stochastic aspect of the simulation is embedded in the probability of a dose-limiting toxicity shown in Figure 4. For example, using the L1 curve, a modeled patient treated at 1.2 mg/kg has a 7% chance of experiencing a dose-limiting toxicity. In other words, for every 100 patients simulated at that dose, on average seven will have a dose-limiting toxicity.

The most important operating characteristics (outputs) of the simulation were:

- The probability that the correct MTD is selected for each dose–toxicity curve, shown in Table 2;
- The mean sample size, across all simulations, required to complete the trial, shown in Table 3.

For four of the six dose–toxicity curves (S, H 1, L1 and L2), the NCRM designs were more likely than the 3+3 design to select the correct MTD. For curve F, both NCRM-1 and NCRM-2 outperformed the 3+3 design. However, for curve H2, the 3+3 design was the most likely to select the correct MTD.

The mean simulated sample size needed to complete the trial ranged from 11.8 to 24.2, with NCRM-1 and NCRM-3 requiring fewer patients than the 3+3 design for all dose–toxicity curves except H2. NCRM-2 required the largest sample size in several scenarios because the design specified that three cohorts must be dosed at MTD, compared with two cohorts in NCRM-1 and NCRM-3.

Overall, the NCRM designs outperformed the 3+3 design for five of the six dose–toxicity curves. The exception was H2, the steepest curve with the highest rate of toxicity but also the one with the least likelihood of occurrence because the drug was thought to be relatively nontoxic in the dosing range. The most accurate of the NCRM designs was NCRM-2, but the improved performance required a larger sample size.

Table 1. Parameters for Neuenschwander continual reassessment method designs.

Parameter	NCRM-1	NCRM-2	NCRM-3
Number of cohorts that must be treated at the maximum tolerated dose before study can be stopped early	Two	Three	Two
Maximum percentage of simulations that can report a probability of dose-limiting toxicity >33% at the maximum tolerated dose (controls the risk of overdosing)	25%	25%	15%

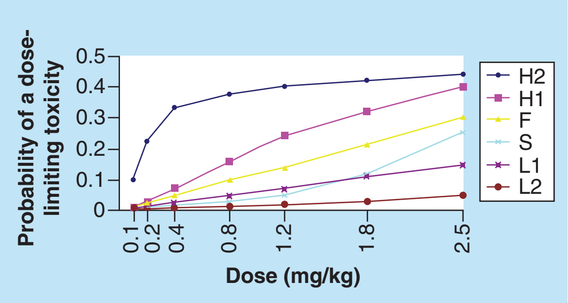

Figure 4. Simulations were conducted on six potential dose–toxicity curves.
F: Flat slope at maximum tolerated dose; H1: High-toxicity profile 1; H2: High-toxicity profile 2; L1: Low-toxicity profile 1; L2: Low-toxicity profile 2; S: Steep slope at maximum tolerated dose.

While failure to identify the MTD can occur with any study design, 3+3 designs carry an increased risk that this incorrect dose is lower—and sometimes substantially lower—than the true MTD. Clearly, this result can lead to lower rates of efficacy in later trials and put further development of the compound at risk. For that reason, it can be useful to investigate the performance of different designs by not only calculating the probability that the incorrect MTD is chosen but also by investigating the distance between the incorrectly selected MTD and the actual MTD. Simulation results suggest that when incorrect MTD is selected, NCRM's selection tends to be close to actual MTD, while 3+3 often significantly underestimates MTD.

Phase II case study: simulations to better define dose–response relationships

The goal of most Phase II studies is to gain a thorough understanding of a drug's dose–response relationship. Phase II studies often use fixed, parallel designs with no more than four equally randomized groups: three dose levels and a control. However, this approach may be inadequate to characterize the limits of the dose–response curve and can result in the choice of suboptimal doses for Phase III evaluation.

In this Phase II learning study, the sponsor wanted to explore design options for evaluation of a new antihypertensive drug. The measured response was the decrease from baseline in diastolic blood pressure, and the smallest clinically relevant difference was defined as ≥10 mm Hg.

The goal of the trial was to identify a dose that would meet two prespecified criteria:

- That the dose is the ED_{95}, defined as the lowest dose that produced a response ≥95% of the estimated maximum response (effective dose);
- That there is a ≥60% chance that the true difference versus placebo exceeds the clinically relevant difference in the potential Phase III patient population.

Mathematically, as long as the model is estimable, the trial will always determine an ED_{95}.

The key question then becomes: 'What is the probability that the mean response at ED_{95} will meet or exceed the specified clinically relevant difference if that

Table 2. Probability that the correct maximum tolerated dose is selected for each dose–toxicity. curve.

Design	Dose–toxicity curves					
	F	S	H1	H2	L1	L2
3+3	0.26	0.52	0.26	0.34	0.66	0.96
NCRM-1	0.27	0.74	0.31	0.31	0.87	0.99
NCRM-2	0.28	0.79	0.36	0.31	0.87	>0.99
NRCM-3	0.16	0.79	0.32	0.27	0.84	0.99

F: Flat slope at maximum tolerated dose; H1: High-toxicity profile 1; H2: High-toxicity profile 2; L1: Low-toxicity profile 1; L2: Low-toxicity profile 2; NCRM: Neuenschwander continual reassessment method; S: Steep slope at maximum tolerated dose.

Table 3. Mean sample size required to complete the trial (Phase I case study).

Design	Dose–toxicity curves					
	F	S	H1	H2	L1	L2
3+3	22.0	23.0	19.7	11.8	22.9	21.9
NCRM-1	17.8	16.4	18.6	15.6	15.5	13.8
NCRM-2	23.2	20.6	24.2	20.3	19.5	16.8
NRCM-3	19.0	18.0	19.3	14.5	16.8	14.1

F: Flat slope at maximum tolerated dose; H1: High-toxicity profile 1; H2: High-toxicity profile 2; L1: Low-toxicity profile 1; L2: Low-toxicity profile 2; NCRM: Neuenschwander continual reassessment method; S: Steep slope at maximum tolerated dose.

dose is used in a Phase III trial?” If the probability is ≥60%, then the Phase II trial met its goal (according to the second criteria) and can be called a success. If the probability is <60%, then the Phase II trial failed to meet its objective and no suitable dose for Phase III trials would be identified. Note that all the numerical thresholds in this case study are subjective and depend on the drug, therapeutic area and end point.

Clinical trial simulation was used to estimate the probability of success for five different study designs, as summarized in Table 4. All of these designs used seven active dose levels and a control (eight arms in total).

Design A was the base case. It was a fixed design with an equal allocation ratio and a randomized sample size of 640 patients; expected overall study duration was 10 months. Responses at a given dose level were assumed to be normally distributed and were analyzed individually against the control (i.e., ignoring any potential dose–response relationship).

Designs B–E included at least one of the following four additional design elements:

- Dose–response model: the mathematical equations used to describe dose–response relationship. In the base case (Design A, no model), each dosing level was analyzed independently. With a dose–response model, results at one dose level help estimations at other dose levels by improving the model;

- Response-adaptive randomization: the scheme in which the probability of a patient being assigned to a specific treatment group is adjusted by comparative analyses of the accumulated outcome responses of previously enrolled patients [14]. In this case study, the first 40 patients were randomized equally among the seven dosing levels and the control group (‘run-in’ or ‘burn-in’ period). Randomization ratios for subsequent patients were based on the outcomes of these initial 40 patients, with more patients assigned to successful doses – that is, doses that had a higher probability of yielding results that were greater than the predefined clinically relevant difference of 10 mm Hg. The only exception was the control arm, which had to include at least 15% of patients (close to the one-in-eight that would be assigned to the control in a study with equal randomization). Randomization ratios were recalculated at four predetermined patient recruitment thresholds;

- Arm dropping: the discontinuation of one or more dose levels when data indicate that dose is not effective. In this case study, the threshold for a low likelihood of success was chosen to be a <10% chance that the mean response at that dose would meet or exceed the predefined clinically relevant difference of 10 mm Hg. A maximum of three arms could be dropped from the study;

- Early stopping criteria for futility or success: the entire trial can be stopped if it is clear that the answer has been reached. In this case study, the trial could be stopped for futility if there was a <30% chance that the mean response at the calculated ED_{95} (the lowest dose that produced a response ≥95% of the estimated maximum response) would meet or exceed the predefined clinically relevant difference. Conversely, the trial could also be stopped for success if there was a >80% chance that the mean response at ED_{95} would meet or exceed the predefined clinically relevant difference. The minimum sample size was set to 250 if stopping early for futility, and to 125 if stopping early for success.

Because the ‘true’ shape of the dose–response curve is unknown at the beginning of the Phase II trial, each of the five study designs was simulated using four different potential dose–response curves (Figure 5): null (no dose-dependent change in response), C1 (response at highest dose was equal to the predefined clinically relevant difference of 10 mm Hg), C2 (response at

Table 4. Study designs for simulation.

Design	Design options			
	Dose–response model	Response adaptive randomization	Arm dropping	Early stopping criteria†
Design A	No	No	No	No
Design B	Yes	No	No	No
Design C	Yes	Yes	No	No
Design D	Yes	Yes	Yes	No
Design E	Yes	Yes	No	Yes

†For futility or success.

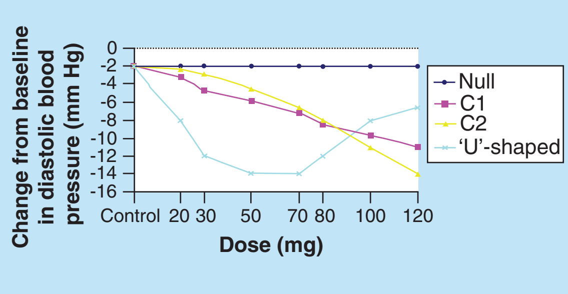

Figure 5. Simulations were conducted on four potential dose–response curves.
C1: Response at highest dose 10 mm Hg; C2: Response at highest dose >10 mm Hg.

highest dose was greater than the clinically relevant difference), and 'U'-shaped (maximum response at middle doses was greater than the clinically relevant difference).

The key operating characteristic compared across the study designs was the probability of success (Table 5). Recall that success was defined as ≥60% chance that the true underlying response at ED_{95} would meet or exceed the clinically relevant difference of 10 mm Hg. The null curve is used to evaluate each design's ability to control type I error (i.e., the probability of identifying an acceptable dose when none exists because the null curve has the same response for all doses).

Designs C and E have much lower probabilities of success for the null curve, therefore controlling type I error rate at the desired 5% level. In addition, designs C and E show the highest rates of success for the 'U'-shaped curve and perform as well as the other designs for the C2 dose–response curve, but they have a slightly lower rate of success for the C1 dose–response curve.

These results demonstrate that, in many cases, adaptive randomization (Design C) and early stopping for futility or success (Design E) can increase the power of a trial while simultaneously controlling the type I error rate. However, based on the probability of success, no single design stands out as being the best under all possible scenarios. Other factors – such as sample size, operational complexity and which simulated dose–response curve is most likely to be correct – need to be considered before making a decision.

For designs A, B and C, the sample size was fixed at 640 patients. For designs D and E, which included arm dropping and early stopping criteria, respectively, the sample size could be adjusted (Table 6). Note that these values indicate the mean sample size, across all simulations, required to complete the trial. The actual number of patients needed in a single 'real' clinical trial will vary.

Based on these results, Design E has the best overall combination of features: a probability of success similar to the other designs across all four potential dose–response curves, an acceptable type I error rate and the smallest mean sample size. By quantifying potential outcomes using the various designs, the simulation results enabled to the sponsor to identify the best balance of the benefits of Design E against the added operational and statistical complexity of the different adaptive trial design features.

Table 5. Probability of success.

Design	Dose–response curves			
	Null	C1	C2	U-shaped
Design A	0.15	0.62	0.73	0.83
Design B	0.15	0.61	0.72	0.82
Design C	0.04	0.56	0.73	0.88
Design D	0.13	0.60	0.71	0.82
Design E	0.05	0.58	0.71	0.89

C1: Response at highest dose 10 mm Hg; C2: Response at highest dose >10 mm Hg.

Table 6. Mean sample size required to complete the trial (Phase II case study).

Design	Dose–response curves			
	Null	C1	C2	U-shaped
Designs A/B/C[†]	640	640	640	640
Design D	554	590	585	614
Design E	444	549	520	492

[†] The sample size was fixed for designs A, B and C.
C1: Response at highest dose 10 mm Hg; C2: Response at highest dose >10 mm Hg.

Phase III case study: simulations to guide group sequential designs

During the learning phase, simulation-informed adaptations increase confidence that the most appropriate dose and study population to demonstrate drug safety and efficacy are determined for the confirmatory phase. With the advance of a drug candidate into Phase III evaluation, adaptations focus on achieving the optimal sample size. Options for this are either:

- The framework of 'group sequential designs' to halt a study early once interim analyses have demonstrated efficacy or indicated futility;
- Increasing the sample size via a sample size re-estimation to ensure that the study has the desired power based on interim results rather than on assumptions made beforehand, which is particularly relevant if there is substantial uncertainty associated with these assumptions.

These relatively simple adaptive design options can provide significant cost savings by allowing for fewer subjects and shorter research time lines. In cases where the sample size re-estimation leads to a sample size increase, the value of these design options is in increasing the chances of a successful outcome for the study. A 2013 report by the Tufts Center for the Study of Drug Development estimated that early stopping for futility and sample size re-estimation, applied across the portfolio, could save sponsors between $100 million and $200 million annually [15].

Phase III study in women's health: design challenges

The sponsor was designing a pivotal Phase III trial to test an investigational women's health therapy against an active comparator, using a noninferiority design with 90% power and a one-sided 2.5% significance level. A noninferiority margin of 10% had been agreed with regulatory authorities. The sponsor wanted to allow early stopping if noninferiority could be established with a smaller sample size. Alternatively, if the investigational drug proved inferior, the trial should be stopped to conserve valuable resources. There was uncertainty around the expected response rate for both for the investigational product and the active comparator; this was expected to be between 30 and 40%.

Simulations were therefore performed to allow informed decision making on the following key design questions:

- What stopping rules should be used?
- How many interim analyses should be conducted and when they should take place?
- How the robustness of the design would be affected by various assumed response rates?

The first set of simulations examined the impact of two possible stopping rules for futility, that is, the inability to demonstrate noninferiority. The O'Brian-Fleming rule requires strong evidence for stopping early but has a smaller impact on the maximum sample size. By contrast, the Pocock rule more easily allows for stopping early but tends to lead to a larger maximum sample size. The O'Brien–Fleming rule was used for stopping for success, applied at every interim analysis except the first, since stopping for success at that point was felt to be unrealistic.

Table 7 shows mean sample sizes across all simulations for three study design choices:

- Stopping rule for futility (O'Brien–Fleming or Pocock);
- Number of interim analyses (two or three). Interim analyses were unequally spaced at 50 and

Table 7. Planned sample sizes for various stopping strategies and numbers of interim analyses.

Number of interim analyses	Analysis	O'Brien–Fleming		Pocock	
		30% response rate	40% response rate	30% response rate	40% response rate
Two	Interim 1	465	531	509	582
	Interim 2	697	797	764	872
	Final	929	1062	1018	1163
Three	Interim 1	376	429	416	476
	Interim 2	563	644	624	713
	Interim 3	751	858	832	951
	Final	939	1073	1040	1189

Table 8. Effects of unequal response rate on stopping probabilities.

Response rate		Analysis	P (stop for)	
Control (%)	Test (%)		Success (%)	Futility (%)
40	38	Interim 1	4.8	2.5
		Interim 2	23.1	6.7
		Interim 3	27.5	8.3
		Final	18.7	8.4
		Overall	74.1	25.9
40	35	Interim 1	1.1	9.5
		Interim 2	7.6	19.0
		Interim 3	14.1	18.9
		Final	15.1	14.7
		Overall	37.9	62.1

75% of maximum sample size for two interims, and at 40, 60 and 80% for three interims;

- Response rate (30 or 40%, but assumed to be the same in both treatment groups).

Table 7 shows the O'Brien–Fleming rule to be better than the Pocock rule for futility stopping, since the Pocock rule always requires a larger sample size than O'Brien–Fleming. It also shows that the mean sample size for a design with three interim analyses is essentially the same as that for a design with two interim analyses. This means that if there is a nontrivial chance of stopping the study early, the design with three interim analyses will almost certainly result in a smaller mean sample size. Regarding the question surrounding response rates, the simulation shows, as expected, that the sample sizes based on a 40% response rate are larger than those based on a 30% response rate. This difference was then explored further (Table 8).

Table 9 presents the chance of stopping for success or futility at each interim analysis for the design with two and three interim analyses as well as the average number of study subjects when both the theoretical and actual response rates are equal to 30%.

Traditionally, design decisions regarding the number and timing of interim analyses are made based on a combination of research experience and guess work. Simulations quantify outcomes for various assumptions and scenarios to support evidence-based decisions. Results of this simulation show, for example, that in a study with two interim analyses, there is a 68% chance of stopping the study with a successful outcome (i.e., noninferiority has been established) and a 7.2% chance of stopping for futility at the second interim analysis which, on average, will occur 49.5 weeks after the first subject is randomized.

Due to the difference in timing of the interim analyses, the design with three interim analyses may stop earlier than the one with two. This needs to be balanced with additional operational effort of conducting a third interim analysis and increasing the average sample size by 10 subjects.

Using the design with three interim analyses and using the O'Brien–Fleming stopping rule for efficacy and futility, simulations were run to evaluate the impact of various combinations of assumed and actual response rates on average timing of the interim analyses and stopping probabilities. Results showed that the conservative assumption of a 40% response rate provided good protection regarding the power (probability of success) when the actual response rate is less than 40%. In this scenario, maximum sample sizes are increased by about 12%. Making the more economical assumption of a 30% response rate leads to a slight increase in the chance of stopping for futility. The overall power is maintained well, but chances of stopping early are reduced, leading to slightly larger average sample sizes.

Table 9. Summary statistics for designs using O'Brien–Fleming stopping rules for both success and futility.

Number of interim analyses	Analysis	Mean number of subjects at time of analysis		Mean time of analysis occurring†	P (stop for)	
		Randomized	Complete		Success (%)	Futility (%)
Two	Interim 1	560.2	441.0	34.9	0.0	2.3
	Interim 2	792.8	662.0	49.5	68.0	7.2
	Final	929.0	879.6	63.8	20.5	2.5
Three	Interim 1	471.6	357.0	29.5	0.0	0.1
	Interim 2	659.2	535.0	41.2	44.1	2.7
	Interim 3	846.6	713.0	52.9	29.9	3.7
	Final	939.0	889.4	64.5	14.5	3.9

†Weeks after first subject randomized.

Table 10. Maximum and average sample sizes for the design with three interim analyses.

Maximum		Mean when treatments are equivalent		Mean when treatments are not equivalent	
30% response rate	40% response rate	30% response rate	40% response rate	30% response rate	40% response rate
939	1073	687.1	785.2	541.6	619.0

Table 10 summarizes maximum and mean sample sizes when the experimental and control treatments are equivalent and when they are not equivalent.

Simulations to further refine trial design

Performing simulations is an iterative process: results of initial simulations suggest new questions, which then can be addressed in subsequent simulations to further evaluate preferred design options in greater detail. In this case, the sponsor was satisfied that the O'Brien–Fleming stopping rule was most efficient and concluded that the conservative 40% response rate, with the resulting increase in sample size, was worthwhile insurance. However, since the chance of stopping for success at the first opportunity (i.e., at the second interim) was relatively high, the sponsor also wanted to investigate the option of allowing stopping for success at all interim analysis, including the first one. The design was also updated to allow a delayed response (six weeks after randomization) and a 5% chance of nonevaluability, features which were not considered in the first simulation.

Results of these new simulations indicated that the cost of a design that also permitted stopping for success at the first interim was negligible, requiring only two additional subjects for the two-interim design and one subject for three-interim design. Permitting stopping for success at the first interim had virtually no effect on the overall probability of stopping for success, but slightly reduced the mean sample size. Since the probability of stopping early was still high, the sponsor decided to investigate the effect of bringing the second interim forward from 75% of maximum sample size to 70%.

A final simulation considered the chances of stopping for futility when the experimental and comparator treatments were not equivalent. Simulation results quantified the higher probability of stopping for futility when the investigational product has a lower response rate than the active comparator (Table 8).

Design decisions based on simulation results

The simulation study quantified the operating characteristics for different design options and, therefore, enabled the sponsor to make an evidence-based choice for design elements best aligned with budget, operational feasibility, time lines and other practical considerations, including its robustness to departures from the assumptions underlying the sample size calculation.

The sponsor decided to implement a group sequential design permitting early stopping for success or futility using O'Brien–Fleming stopping boundaries with two interim analyses (when the primary efficacy end point is available from 50 and 70% of the maximum expected number of subjects). Overall, simulations gave the sponsor more confidence in the adaptive design decisions for the study.

A conventional fixed sample size study with the same objectives and basic design parameters as the adaptive design discussed here would require a sample size of 1009 evaluable subjects. The maximum sample size for the adaptive design chosen by the sponsor is 1056 evaluable subjects, an increase of 4.7%. However, the adaptive study's mean sample size is 650 if the test treatment is not equivalent to the control, and 781 if it is equivalent, leading to sample size reductions of 36 and 23%, respectively. The benefits, both in terms of reduced development time and expenditure, are obvious.

Future perspective

Clinical trial simulation provides the means to optimize adaptive trial designs to improve the likelihood of success at each phase of clinical evaluation. Better decisions in learning during Phases I and II reduce the likelihood of costly Phase III failures. Given the ever-increasing pressures on research costs and time lines, the use of simulation-informed adaptive design is likely to expand significantly during the next five years; regulatory guidance will also increase as regulators and industry gain experience with this advancing approach to trial design.

Adaptive designs are used to increase the probability that a clinical trial will efficiently answer the question it was intended to address. The intrinsic complexity of adaptive design makes clinical trial simulation an invaluable tool to quantify the outcomes of various design options, enabling drug developers to make evidence-based choices regarding potential design adaptations – especially in complex situations with multiple interacting parameters. Although successful simulation outcomes do not guarantee a successful trial, they can significantly improve decision making, provided that the assumptions used to build the simulations are valid.

Informed by well-conducted simulations, adaptive designs can, in some cases, reduce sample sizes and research time lines, but they also require additional expertise, time and investment to conduct. Simulations may, in fact, lead sponsors to conclude that more complex, costly designs are more appropriate to answer a given research question.

The great value of simulation-guided adaptive design is that it has the potential to make the best use of available resources by increasing the sponsor's confidence that a specific study will succeed in answering the question it was designed to address. Incremental improvements in learning phases culminate in overall improvement across the entire development program, greatly improving chances of success in confirmatory Phase III studies.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Professional technical writing assistance used in the production of this manuscript was funded by PPD.

Executive summary

- Clinical trial simulation compares the likely outcomes of various design scenarios based on what is already known, or assumed to be known, about the drug. Results are used to inform decision making in the selection of the best options for trial design.
- The intrinsic complexity of adaptive trial designs makes simulation particularly valuable as a means to identify adaptive elements that will optimize the probability of success. Better decisions during learning Phases I and II reduce the likelihood of costly failures in Phase III.
- A Phase I case study illustrates the use of simulations to compare the capabilities of the Neuenschwander continuous reassessment method and the 3+3 method to determine the maximum tolerated dose. Results showed that the Neuenschwander continuous reassessment method design had a smaller average sample size and greater chance of determining the correct maximum tolerated dose under almost any dose–toxicity assumption.
- A Phase II case study illustrates the use of simulations to incorporate adaptive randomization and early stopping for futility or success. Results informed a dose–response study design that could increase the power of the trial, while simultaneously controlling the type I error rate and reducing the average sample size.
- A Phase III case study illustrates the use of simulations to quantify outcomes for various assumptions and scenarios regarding the number and timing of interim analyses and the rules used to allow stopping for futility and success. Results enabled the sponsor to select a design with only a 5% increase in maximum sample size compared with a conventional fixed sample size, but with a 23–36% reduction in average sample size (after taking into account the possibility of early stopping).
- Clinical trial simulation is an iterative process in which results of initial simulations suggest new questions that can be addressed in subsequent simulations to further refine design options. By quantifying operating characteristics for different options, simulations help identify design choices that are best aligned with budget, operational feasibility, time lines and other practical considerations.
- The great value of simulation-guided adaptive design is that it has the potential to make the best use of available resources by increasing the sponsor's confidence that a specific study will succeed in answering the question it was designed to address. Adaptive design usually requires a longer and more complex study initiation phase but is likely to lead to more efficient long-term program development.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest.

1 Morgan S, Grootendorst P, Lexchin J, Cunningham L, Greyson D. The cost of drug development: a systematic review. *Health Policy* 100(1), 4–17 (2011).

2 Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* 32(1), 40–51 (2014).

3 Hay M, Rosenthal J, Thomas D, Craighead J. Bio/BioMedTracker clinical trial success rates study. Presented at: BIO CEO & Investor Conference, 15 February 2011. http://insidebioia.files.wordpress.com

4 Desmond-Hellmann S. The cost of creating a new drug now $5 billion, pushing big pharma to change. Forbes Pharma & Healthcare, September 11, 2013. http://www.forbes.com

5 US Food and Drug Administration. Innovation or stagnation: Critical Path Opportunities Report. www.fda.gov

6 Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat. Med.* 25(24), 4279–4292 (2006)

7 US Food and Drug Administration. Draft guidance for industry: adaptive design clinical trials for drugs and biologics.
www.fda.gov

•• **Detailed regulatory expectations on design and analysis of adaptive trials**

8 PPD. Implementing adaptive trial design: operational considerations and the role of the CRO. PPD White Paper, www.ppdi.com

9 Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A comparison of two phase I trial designs. *Stat. Med.* 13(18), 1799–1806 (1994)

10 Reiner E, Paoletti X, O'Quigley J. Operating characteristics of the standard phase I clinical trial design. *Comput. Stat. Data Anal.* 30(3), 303–315 (1999).

11 Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to Phase I cancer trials. *Stat. Med.* 27(13), 2420–2439 (2008).

12 Ji Y, Wang SJ. Modified toxicity probability interval design: a safer and more reliable method than the 3+3 design for practical Phase I trials. *J. Clin. Oncol.* 31(14), 1785–1791 (2013).

13 Ji Y, Pan H. A comparison of adaptive designs for Phase I clinical trials. In: *Important Considerations for Clinical Trial Methodologies.* Berger VW & Zhang XC (Eds.) Future Science Book Series, London, UK, 102–114 (2013).

14 Fardipour P, Littman G, Burns DD *et al.* Planning and executing response-adaptive learn-phase clinical trials: 1. the process. *Ther. Innov. Regul. Sci.* 43(6), 713–723 (2009).

15 Getz K, Stergiopoulos S, Kim J. The adoption and impact of adaptive trial designs. R&D senior leadership brief. Tufts Center for the Study of Drug Development, Tufts University. http://csdd.tufts.edu

• **Report on practical use and implementation of adaptive design.**