Clinical Cancer Research

The Impact of Early-Phase Trial Design in the Drug Development Process

Mark R. Conaway and Gina R. Petroni

Abstract

Purpose: Many of the therapeutic agents that are being used currently were developed using the 3+3 decision rule for dose finding. Over the past 30 years, several dose-finding designs have been proposed and evaluated, including the "continual reassessment method" (CRM) and the "Bayesian optimal interval design" (BOIN). This research investigates the role of the choice of an early-phase design on the likelihood that drugs entering the drug development pipeline will have 2 successful phase III trials.

Experimental Design: Using simulation, each agent in a population of hypothetical agents was tracked through the drug development process, from initial dose finding to 2 confirmatory phase III trials. Varying the designs of the phase I, II, and III trials allows for an assessment of the effect of the choice of designs on the proportion of agents with successful phase III trials.

Results: The results indicate that using the CRM or BOIN, rather than the 3+3, substantially enhances the proportion of effective agents that have successful phase III trials, with the CRM having a greater effect than BOIN. A larger phase II trial magnifies the effect of the phase I design.

Conclusions: The results underscore the importance of the choice of the early-phase designs. Use of the 3+3 results in fewer agents with successful phase III trials compared with the CRM or BOIN. The difference is more pronounced among highly effective agents. In addition, the results show the importance of a sufficiently powered phase II trial.

Introduction

Numerous articles in the statistical and clinical trial literature have compared the 3+3 decision rule with the continual reassessment method (CRM; refs. 1–6) for phase I dose-finding trials. Other designs have been proposed recently, including the Bayesian optimal interval design (BOIN; refs. 7–11), and these designs have been compared with the 3+3 rule. Consistently, these papers have reported that both the CRM and BOIN are superior to the 3+3 in terms of identifying the MTD and in allocating participants to doses near the MTD.

The main criterion used to make this comparison, the percentage of times the MTD is correctly identified, does not resonate with clinicians and does not address the long-term implications of the choice of the dose finding design. In this article, we consider a different criterion, assessing the effect of the early stage design on the probability that an agent will be shown to be significantly different than the current standard in 2 consecutive phase III trials. The overall goal of this research is to assess how the choice of the early-phase design affects the outcome of the drug development process.

Materials and Methods

A description of the standard and new therapies

We simulate the entire phase I, II, and III process of drug testing starting from the dose-finding trial. To make the simulations realistic, we use published data on dose-limiting toxicities (DLTs), objective response rates (ORRs), and overall survival (OS) for 4 doses of pembrolizumab in participants with non–small cell lung cancer (12–14). We recognize that pembrolizumab is a targeted therapy, and the doses used in the trials did not necessarily represent increasing levels of a single agent. Our goal is not to review the process for this specific drug, but to use a concrete example to evaluate, under realistic situations, the effect of the dose-finding design on the probability that a hypothetical new agent, with an efficacy and toxicity profile similar to pembrolizumab, would be shown to be significantly better than the current standard. We assume that, with the current standard therapy, the toxicity rate is 20%, the ORR is 0.10, and the median survival time is 6 months. The toxicity rate of 20% was chosen to reflect that most agents in use currently were developed using the 3+3 design, which targets a DLT rate between 10% and 30% (15, 16). Four doses of an agent entering the development pipeline are under consideration. These doses have DLT probabilities equal to 0.05, 0.12, 0.15, and 0.20, ORRs of 0.15, 0.25, 0.35, and 0.40, and median survival times of 6.25, 6.50, 7.50, and 10.0 months, respectively.

The simulations address how the effectiveness of the agent affects the probability that a new agent is shown to be significantly better than the current standard. The toxicity, response, and median survival profiles for these agents are shown in Table 1, indexed by a parameter, Δ, which ranges from 0 to 1.5. The example in the previous paragraph corresponds to $\Delta = 1$. A value of $\Delta = 0$ means that for each dose of the new agent, the ORR and the median OS are the same as the current standard. A value of $\Delta = 1.5$ means that the new agent is more effective at all doses than

Division of Translational Research and Applied Statistics, Department of Public Health Sciences, University of Virginia Health System, Charlottesville, Virginia.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Mark R. Conaway, University of Virginia, Hospital West 3181, Charlottesville, VA 22908. Phone: 434-924-8510; Fax: 434-243-5787; E-mail: mrc6j@Virginia.EDU

doi: 10.1158/1078-0432.CCR-18-0203

Translational Relevance

Every new agent developed in the laboratory is required to go through the drug development process for approval before it can be used to treat patients in the general population. This article studies components of that process, assessing how the choice of designs at each stage in the process affects whether or not even a highly effective agent entering the drug development pipeline will be observed to have 2 consecutive successful phase III trials.

the example in the previous paragraph, with ORRs of 0.175, 0.325, 0.400, and 0.625 and median survival times of 6.375, 6.75, 8.25, and 12.0 months, respectively, at the 4 dose levels.

The drug development process

Our simulation of the process begins with a dose-finding study to find the MTD based on a toxicity endpoint, commonly the DLT. For the 3+3, the MTD is the highest dose with fewer than 2 of 6 participants who experience a DLT. For the CRM, the MTD is defined as the dose with the DLT probability closest to the prespecified target toxicity rate. We are using the 2-stage likelihood CRM (17), consisting of a rule-based run-in stage and a modeling stage. In actual trial settings, we most commonly use cohorts of size 1 in the modeling phase of the CRM, but at the request of the senior editor, we are using cohorts of size 2 in the rule-based and modeling stages of the CRM simulations. The working model for the DLT probabilities in the CRM is guided by the recommendations of Lee and Cheung (18). We also evaluate another recently proposed design, the BOIN with cohorts of size 2 for comparisons with the 3+3. All of the dose-finding designs have stopping rules for safety based on the participants assigned to the lowest dose level who experience a DLT. In the 3+3, if de-escalation is indicated at the lowest dose level, the study is stopped. For the CRM, we compute the lower bound of a 1-sided 90% confidence interval (19) for the DLT probability at the lowest dose level. If the lower bound exceeds the target, the study is stopped. With BOIN, the stopping rule is based on the posterior probability that the DLT probability at the lowest dose exceeds the target DLT threshold. If the posterior probability is sufficiently large, the study is stopped. For all designs, if the stopping rule is activated, all of the doses are considered too toxic and the drug will not be considered for further study.

In addition, the 3+3 may include an expansion cohort in which an additional set of participants are enrolled on the dose identified as the MTD. This is the most difficult step in the process to simulate, because in practice, there are rarely formal rules applied to the data collected in the dose-expansion cohort (20). In our simulations, in designs that use an expansion cohort, if there are too few responders in the expansion cohort, the agent at the identified dose is not considered for further testing. The expansion cohort can also be used to assess toxicity, and to modify the dose identified in the previous dose-finding phase. Given a target level of toxicity, assumed to be 20% in our simulations, the recommended dose for the phase II is reduced by 1 level if the number of participants experiencing DLTs in the expansion cohort is significantly greater than the target, using a 1-sample binomial test. When the expansion cohort is conducted at the lowest dose level, and too much toxicity is observed, the agent is not considered for further testing.

Following the dose-finding design, a 2-stage Simon design (21) is conducted using an objective response endpoint at the dose identified as the MTD. With a sufficient number of observed responses after each stage of the Simon design, the new agent is carried forward at the chosen dose into a randomized comparative trial with OS as the outcome and the log-rank test is used for testing significance. In addition, in the phase III trial, at the end of the trial, we computed a 1-sided binomial test for differences in toxicity proportions between the current standard and a new agent, rejecting the null hypothesis of equal toxicity when the new agent showed significantly greater toxicity than the current standard. A design that tended to over-estimate the MTD would be penalized at this stage, because there would be a greater likelihood of rejecting equality of toxicity proportions at the phase III trial. We added a confirmatory phase III trial, with the same sample size as the first phase III trial, conducted only if the first phase III trial demonstrated a significant difference in survival time between the new agent and control, and with no significant increase in toxicity. With these decision rules, there are 9 possible outcomes for the drug development process.

1. The dose-finding trial finds that all dose levels under consideration are too toxic.
2. The dose expansion cohort has too few responders.
3. Too much toxicity is observed in the dose expansion cohort conducted at the lowest dose.
4. Too few responders are observed in stage 1 of the Simon design.
5. Too few responders are observed at the end of stage 2 of the Simon design.
6. The initial phase III trial shows significantly greater toxicity with the new agent.
7. The initial phase III trial shows no significant difference in OS.
8. The initial phase III trial shows a significant difference in OS with no significant increase in toxicity, but the confirmatory trial either shows no significant difference in survival time, or shows a significant increase in toxicity.
9. Both the initial and confirmatory phase III trials show a significant difference in OS with no significant increase in toxicity.

Table 1. Assumed toxicity, ORR, and median OS for a range of profiles for a hypothetical new agent at 4 dose levels

Treatment	DLT proportion	ORR	Median OS (mo)
Current standard		0.10	6
New agent: Dose level 1	0.05	0.10 + Δ (0.15–0.10)	6.0 + Δ (6.25–6.0)
New agent: Dose level 2	0.12	0.10 + Δ (0.25–0.10)	6.0 + Δ (6.50–6.0)
New agent: Dose level 3	0.15	0.10 + Δ (0.30–0.10)	6.0 + Δ (7.50–6.0)
New agent: Dose level 4	0.20	0.10 + Δ (0.45–0.10)	6.0 + Δ (10.0–6.0)

Details of the phase I, II, and III designs

For the phase I design, we will consider 6 possibilities:

1. A 3+3 with an expansion cohort of 12 participants (3+3, EC12).
2. A 3+3 with an expansion cohort of 20 participants (3+3, EC20).
3. The CRM with 24 participants, cohorts of size 2, and no expansion cohort (CRM, $n = 24$).
4. The CRM with 36 participants, cohorts of size 2, and no expansion cohort (CRM, $n = 36$).
5. BOIN with 24 participants, cohorts of size 2, and no expansion cohort (BOIN, $n = 24$).
6. BOIN with 36 participants, cohorts of size 2, and no expansion cohort (BOIN, $n = 36$).

There is no expansion cohort used in our simulations of the CRM or BOIN, but we allow for a preliminary check for response in the simulated phase I trial, using the same percentage allowed for the expansion cohort with the 3+3. To move the agent into phase II testing, we require at least 2 responders among all 24 participants, or at least 3 responders in 36 participants in the CRM or BOIN designs. Although this is generally not part of CRM or BOIN, which are based on toxicity only, we believe that sponsors are reluctant to move forward with an agent without a preliminary indication of efficacy (22).

For the phase II design, we use 2 versions of the Simon optimal design each with a null response rate of 0.10. The designs differ in the alternative hypothesis response rate, 0.30 or 0.25. For an alternative response rate of 0.30, the first stage of the design enrolls 10 participants. If 1 or fewer of these participants respond, the trial is stopped and the agent is not considered for further testing. If 2 or more first-stage participants respond, an additional 19 participants are enrolled. If 6 or more of the 29 participants in total are responders, the agent proceeds to phase III testing. For an alternative response rate of 0.25, the first stage of the design enrolls 18 participants. If 2 or fewer of these participants respond, the trial is stopped and the agent is not considered for further testing. If 3 or more first-stage participants respond, an additional 25 participants are enrolled. If 8 or more of the 43 participants in total respond, the agent proceeds to phase III testing.

For the phase III trial, we use an accrual rate of 23.5 participants per month, similar to the accrual rate in Herbst and colleagues (12), with an accrual period of 12.6 months and a follow-up period of 6 months. Generated survival times are exponentially distributed. Accrual was uniformly distributed over the 12.6-month accrual period; participants are censored if they were still alive at 18.6 months. The sample size for the phase III trial is based on comparing median OS of 6 months with the current standard versus 9 months for the new agent. Enrolling 296 participants over a 12.6-month accrual period yields 80% power for the log-rank test, with a 2-sided significance level of 5%. For an alternative median survival of 8 months with the new agent, the sample size is 498 participants accrued over a 21.2-month accrual period. Participants were censored if still alive at 27.2 months.

Generating populations of new agents

Another way to evaluate the drug development process is to consider a population of new agents, each with its own toxicity, objective response, and median survival profiles. We generate populations of 20,000 new agents, and for each agent in the population, we simulate a single set of phase I–II–III trials. We evaluate the effects of the designs of the phase I–II–III studies by estimating the proportion of the population of agents for which 2 successful phase III trials are observed.

The model for generating the toxicity, response, and median survival profiles in the population of new agents is based on an S-shaped curve for toxicity, response, and median survival,

$$f(x) = \beta + \frac{\alpha - \beta}{1 + e^{-\kappa(logit(x))}}$$

where x is evaluated on a grid from 0.001 to 0.999 in increments of 0.001, $\kappa > 0$, and $\alpha > \beta$. With this model, β is the minimum toxicity, response rate, or median survival level for a drug at any dose, and α is the maximum toxicity, response rate, or median survival for the drug at any dose. For each drug attribute of toxicity, response rate, and median survival, we generate α, β, and κ to create drug profiles that were not too toxic, and had ORRs and median survival that are at least as great as those of the current standard, which has an assumed toxicity rate of 20%, an ORR of 0.10, and a median survival of 6 months. Even though the toxicity, response, and survival profiles are generated from a model with the same functional form, this form is sufficiently flexible to encompass a variety of shapes, and allow for cases where the toxicity, response, and survival are unrelated to dose.

Most dose-finding trials are done with discrete doses. For each agent in the population, we randomly choose an integer number of dose levels to be tested in the phase I trial, ranging from 3 to 8, with probabilities 30%, 25%, 20%, 12.5%, 5%, and 2.5%, meaning that on average, we expect 30% of trials to have 3 dose levels, 25% of trials to have 4 dose levels, and only 2.5% to have as many as 8 dose levels. Once the number of levels, n^*, was chosen, we randomly choose n^* sorted values of x from the 999 grid values that range from 0.001 to 0.999. To ensure spacing between the dose levels, the random selection chose 1 x value from each of the n^* intervals, $[(i\text{-}1)^*k, i^*k]$, where $i = 1, 2, \ldots, n^*$ and $k = 999/n^*$.

Simulating the drug development process on the population of agents

The sample size for the CRM and BOIN is chosen to equal 6 times the number of doses under consideration, equal to the maximum number that could be used in the 3+3. An expansion cohort of size 20 is used with the 3+3. The same phase II and phase III trial designs described in the previous section are used. Each of the agents in the population enters the drug development pipeline, and the frequency of the 9 possible outcomes listed in section 2.2 was recorded over the population of agents. This process was repeated 10 times for each population, and we report the proportion of agents with 2 successful phase III trials, averaged over the 10 repetitions.

A population that matches the agents entering the pipeline from 1993 to 2004

DiMasi and colleagues (23) tabulated the results of the drug development process for 625 cancer drugs that entered the drug development pipeline between 1993 and 2004. Overall, 75% of agents successfully made the transition from phase I to phase II testing. Of the agents entering phase II testing, 42% were subsequently tested in a phase III trial. Overall, 13% of agents went from phase I testing to approval. These transition proportions

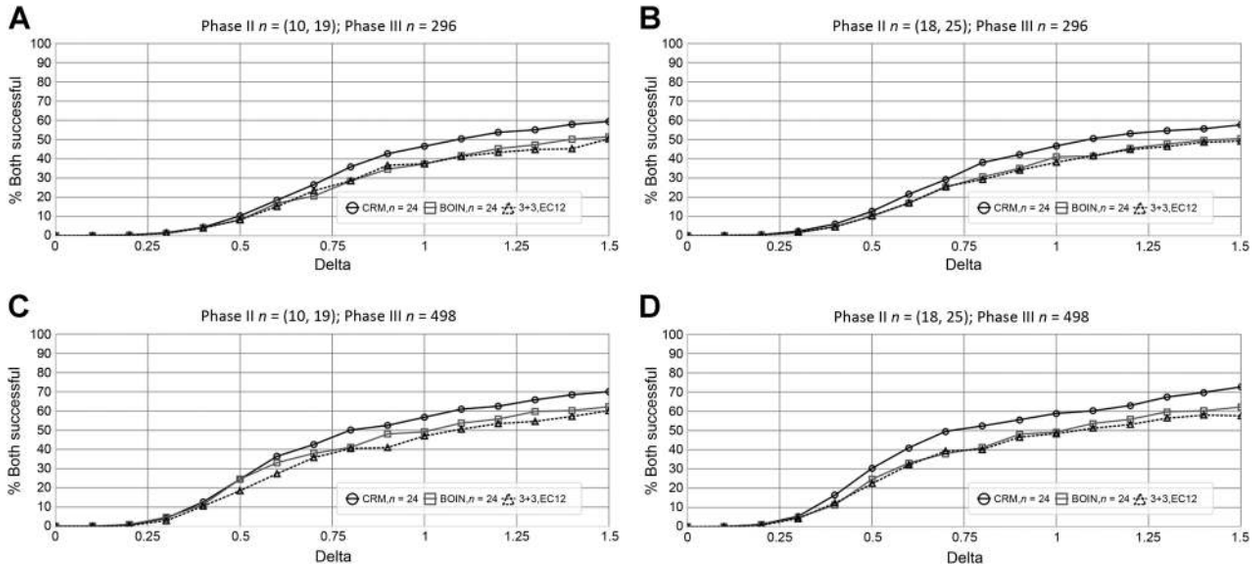

Figure 1.

A–D, Probability of a significant phase III trial for different dose-finding designs by increasing levels of efficacy relative to the standard treatment for CRM and BOIN with sample size $n = 24$ and 3+3 with an expansion cohort of size 12.

guide the parameters chosen to create a population of new agents. Supplementary Table S1 displays the values for the parameters we use to generate the toxicity, response, and median survival profiles for this population. Figure S1 in the Supplementary Material displays 100 toxicity, response rate, and median survival curves randomly chosen from the 20,000 generated drug profiles; Fig. S2 in the Supplementary Material displays the toxicity, response rate, and median survivals for randomly chosen agents with 4, 6, or 8 dose levels in the phase I trial.

We simulated the transition probabilities with the population generated from these parameters, assuming that the process is a 3+3 with an expansion cohort of 20 participants, a Simon phase II design with stage sample sizes (10, 19), and a phase III trial of 296 participants. With this process, 75% of 20,000 agents in the generated population transitioned from phase I to phase II; 43% of agents tested in phase II were subsequently tested in phase III. Overall, if we require 2 significant phase III trials, with no significant increase in toxicity, for "approval," then 14% of

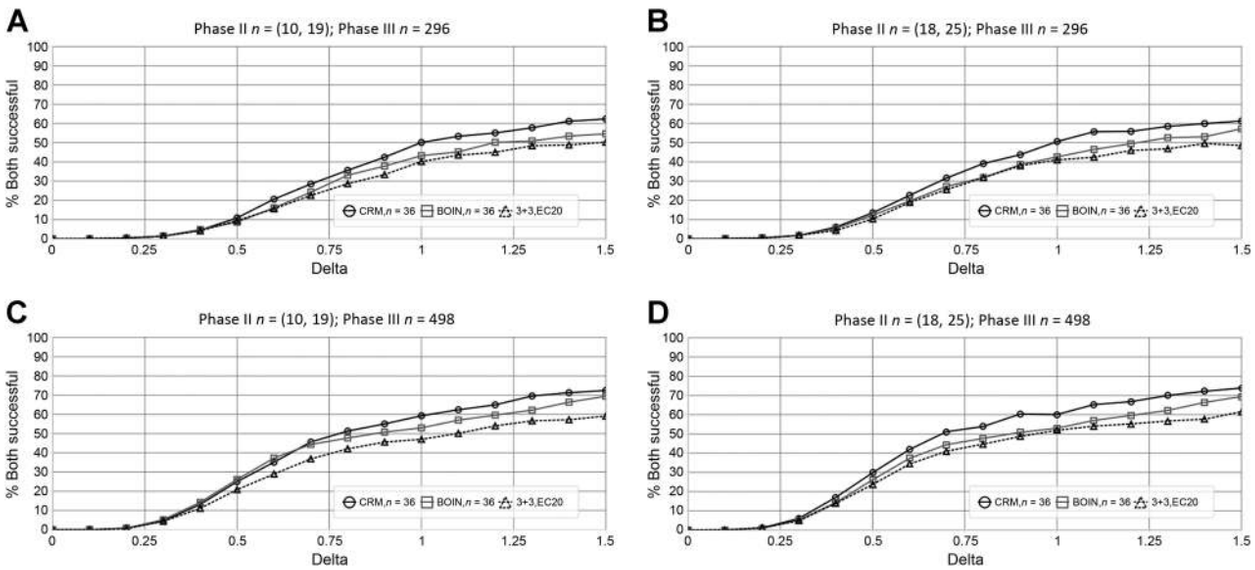

Figure 2.

A–D, Probability of a significant phase III trial for different dose-finding designs by increasing levels of efficacy relative to the standard treatment for CRM and BOIN with sample size $n = 36$ and 3+3 with an expansion cohort of size 20.

agents in this population would go from phase I to approval. We recognize that different phase I, II, and III designs were used for the 625 drugs entering the pipeline between 1994 and 2003, but these results suggest that the population of agents is a realistic model for agents entering the development process.

A population of agents with low toxicity

In an era of targeted and biological agents, we can anticipate agents with lower toxicity than were considered in the previous section. To assess the role of early stage designs in this situation, we generate a population of agents with the same parameters for response and median survival as in the previous section, but with agents with lower toxicity profiles. This population is one for which the effect of dose finding could be expected to be smaller than for other populations. For agents with low toxicity and robust median survival profiles relative to the standard, even a suboptimal dose selection would still put an agent into phase II and phase III testing that has low toxicity and conveys a survival advantage over the current standard.

Results

Results for a single drug, $\Delta = 1$, in Table 1

With 6 phase I, 2 phase II, and 2 phase III designs under consideration, there are a total of $6 \times 2 \times 2 = 24$ phase I–II–III processes to be evaluated. Table 2 shows summaries of the 9 possible trial outcomes, based on 2,500 simulated drug development processes, for each of the 24 combinations, using the toxicity, ORR, and OS values in the column with $\Delta = 1$ in Table 1. The results suggest that the CRM design is associated with an increase in the probability of a successful phase III trial compared with the 3+3, even with an expansion cohort of size 20. Averaged over the 4 phase II and phase III designs, even a small CRM trial ($n = 24$) is associated with a 7.7 percentage point increase in the probability of 2 successful phase III trials. Using a larger CRM ($n = 36$) adds another 4.4 percentage points to the probability of success, for an average of a 12.2 percentage point increase in the probability of 2 successful phase III trials. BOIN, with a sample size of 24, is comparable with the 3+3 with an expansion cohort of 20 patients, but shows an increase of approximately 5 percentage points compared with the 3+3.

Figure 1 displays the probability that both the initial and confirmatory phase III trials show a significant benefit, with no significant differences in toxicity, for each of the 24 combinations of 6 phase I, 2 phase II, and 2 phase III designs as a function of the index Δ. With $n = 24$ participants, the CRM dominates the 3+3. There is little separation between the methods for small values of Δ, which are agents that do not differ much from the current standard. The separation becomes greater with increasing values of Δ, suggesting that the greater the benefit of the new agent relative to the current standard, the greater the benefit in using the CRM over the 3+3. For $n = 36$, BOIN shows a substantial advantage over the 3+3, but the effect is not as great as that observed for the CRM (Fig. 2).

Results for a population that matches the agents entering the pipeline from 1993 to 2004

For 2,775 (14%) of the generated profiles, as described in section 2.6., all of the dose levels under consideration in the phase I trial have a toxicity probability greater than 20%. For the 17,225 drugs with at least 1 safe dose, Fig. 3 shows the average

Table 2. Results of 2,500 simulated phase I–II–III trials

Phase I	Phase II sample size	Phase III sample size	Stop after phase I[a]	Stop after phase II[b]	Initial phase III not successful[c]	Initial phase III successful, second phase III not successful[d]	Both phase III trials successful[e]
3+3, EC12	(10, 19)	296	5.2	15.1	30.3	10.6	38.8
3+3, EC20	(10, 19)	296	5.4	14.2	30.9	10.4	39.0
BOIN, $n = 24$	(10, 19)	296	1.6	11.1	38.4	11.5	37.3
BOIN, $n = 36$	(10, 19)	296	1.1	9.6	34.5	11.6	43.2
CRM, $n = 24$	(10, 19)	296	0.8	8.3	32.4	12.0	46.5
CRM, $n = 36$	(10, 19)	296	0.4	7.4	31.5	10.6	50.1
3+3, EC12	(18, 25)	296	4.8	10.5	34.9	11.2	38.5
3+3, EC20	(18, 25)	296	4.4	9.9	35.6	11.3	38.8
BOIN, $n = 24$	(18, 25)	296	1.4	5.5	40.3	11.8	41.0
BOIN, $n = 36$	(18, 25)	296	1.7	4.4	39.8	11.5	42.6
CRM, $n = 24$	(18, 25)	296	1.0	3.9	36.6	11.8	46.7
CRM, $n = 36$	(18, 25)	296	0.5	2.5	34.2	12.2	50.6
3+3, EC12	(10, 19)	498	5.6	14.0	25.0	8.6	46.8
3+3, EC20	(10, 19)	498	4.9	14.9	20.8	9.0	50.4
BOIN, $n = 24$	(10, 19)	498	2.0	10.7	29.2	10.1	48.0
BOIN, $n = 36$	(10, 19)	498	1.9	11.0	23.7	10.4	53.0
CRM, $n = 24$	(10, 19)	498	1.0	8.7	22.7	10.8	56.8
CRM, $n = 36$	(10, 19)	498	0.2	7.4	22.7	10.4	59.3
3+3, EC12	(18, 25)	498	5.6	8.8	27.9	10.5	47.2
3+3, EC20	(18, 25)	498	4.8	9.4	26.2	9.4	50.2
BOIN, $n = 24$	(18, 25)	498	1.9	5.8	31.8	11.2	49.2
BOIN, $n = 36$	(18, 25)	498	1.6	4.3	29.6	11.5	53.0
CRM, $n = 24$	(18, 25)	498	0.8	4.1	26.0	10.2	58.9
CRM, $n = 36$	(18, 25)	498	0.3	2.7	25.0	11.8	60.1

[a]Includes trial outcomes 1 to 3: all doses too toxic in dose-finding or expansion cohort, or too few responders.
[b]Includes trial outcomes 4 and 5: too few responders in either stage 1 or stage 2 of Simon phase II design.
[c]Includes trial outcomes 6 and 7: initial phase I trial not successful due to excessive toxicity or insufficient efficacy.
[d]Trial outcome 8: the initial phase III trial was successful, but the confirmatory phase III trial showed excessive toxicity or insufficient efficacy.
[e]Trial outcome 9: both phase III trials successful.

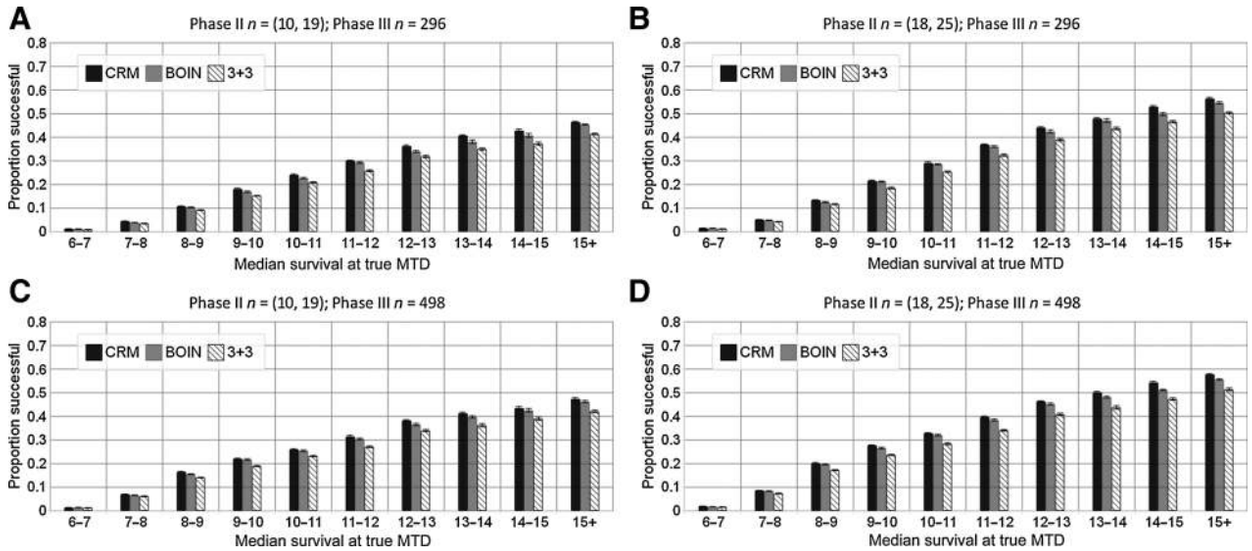

Figure 3.

A–D, Proportion of agents with 2 successful phase III trials among 17,225 drugs (section 3.2) with at least 1 safe dose by true median survival at the true MTD.

proportion of agents with 2 successful phase III trials; these proportions are tabulated in Supplementary Table S2. The horizontal axis is the median survival, grouped by month, at the true MTD, defined as the highest dose under consideration with a DLT probability of 20% or less. The results mirror those presented earlier in this article: the proportion of effective agents shown to be significantly better than the current standard in 2 successive phase III trials is greater with the CRM than with the 3+3. The difference increases with increasing effectiveness of the drug relative to the current standard. Overall, the proportions of agents with 2 successful phase III trials using the 4 combinations of phase II–III designs are 2.3, 2.5, 2.4, and 3.2 percentage points greater with the CRM than the 3+3 and 1.5, 1.8, 2.0, and 2.3 percentage points greater with BOIN than the 3+3. The overall differences are dampened by the number of relatively ineffective therapies in the population. Of the 17,225 drugs, 3,631 (21%) have median survival times between 6 and 7 months, and only 1.3%, 1.4%, and 1.5% of agents have 2 successful phase III trials with the 3+3, CRM, and BOIN, respectively, averaging over the four phase II–III designs. Among agents with a median survival of at least 12 months, a 100% increase over the current standard, the proportions are 5.1, 5.6, 4.8, and 6.2 percentage points greater with the CRM and 3.3, 3.7, 3.5, and 4.2 percentage points greater with BOIN than the 3+3 with an expansion cohort of 20. Averaging over the 4 possible phase II and phase III designs, the proportion of agents with 2 successful phase III trials is 2.6 percentage points greater with the CRM or 1.9 percentage points greater with BOIN than with the 3+3, indicating that using the CRM may have produced an additional 16 to 17 and BOIN an additional 11 to 12 approved agents among the 625 agents entering the pipeline from 1993 to 2004. Comparing panels A with B and panels C with D, in Fig. 3, indicates that a larger phase II trial also has a substantial effect on the probability of success.

In many ways, this example underestimates the limitations of the 3+3 relative to the CRM or BOIN. In the simulations, both the CRM and BOIN used cohorts of size 2; better performance is achieved with cohorts of size 1, as originally proposed in the CRM. In addition, the target toxicity for the CRM and BOIN was set to 20%. This level was chosen to match the 3+3, which has been shown to identify the point on the dose toxicity curve with a 15% to 30% toxicity rate (1, 15–16) not the often-perceived 33rd percentile suggested by the decision rules. One of the additional advantages of the CRM or BOIN is flexibility in setting the target toxicity threshold, for example, increasing the threshold in cases where the main DLTs might be transient or easily reversible. Allowing for a greater threshold for toxicity would result in choosing doses with greater ORRs and longer median survival. Results presented in Supplementary Table S4 and Supplementary Fig. S3 indicate that this is the case. If the target toxicity rate was set at 30%, then overall, the percentage of effective therapies with 2 successful phase III trials is 10.3 percentage points greater with the CRM and 9.8 percentage points greater with BOIN, than with the 3+3.

Results for a population of agents with low toxicity

The results for the proportion of agents, generated as in section 2.7, with 2 successful phase III trials are shown in Fig. 4. As in the earlier populations, the use of the CRM for dose finding is associated with a greater proportion of agents proceeding through 2 successful phase III trials. Overall, the proportions of agents with 2 successful phase III trials using the 4 combinations of phase II and phase III designs are 3.3, 3.8, 3.7, and 4.0, averaging 3.7 percentage points greater with the CRM, than the 3+3 and 2.4, 2.8, 3.0, and 3.2, averaging 2.9 percentage points greater with BOIN than the 3+3. Among agents with a median survival at least 12 months, the increase in the proportions over the 3+3 for the CRM are 5.3, 6.1, 5.2, and 5.5, an average of 5.5 percentage points. The increase in the proportions for BOIN over the 3+3 are 3.7, 4.7, 4.0, and 4.3, an average of 4.2 percentage points. Similar to Fig. 3, Fig. 4 also highlights the importance of a well-powered phase II trial.

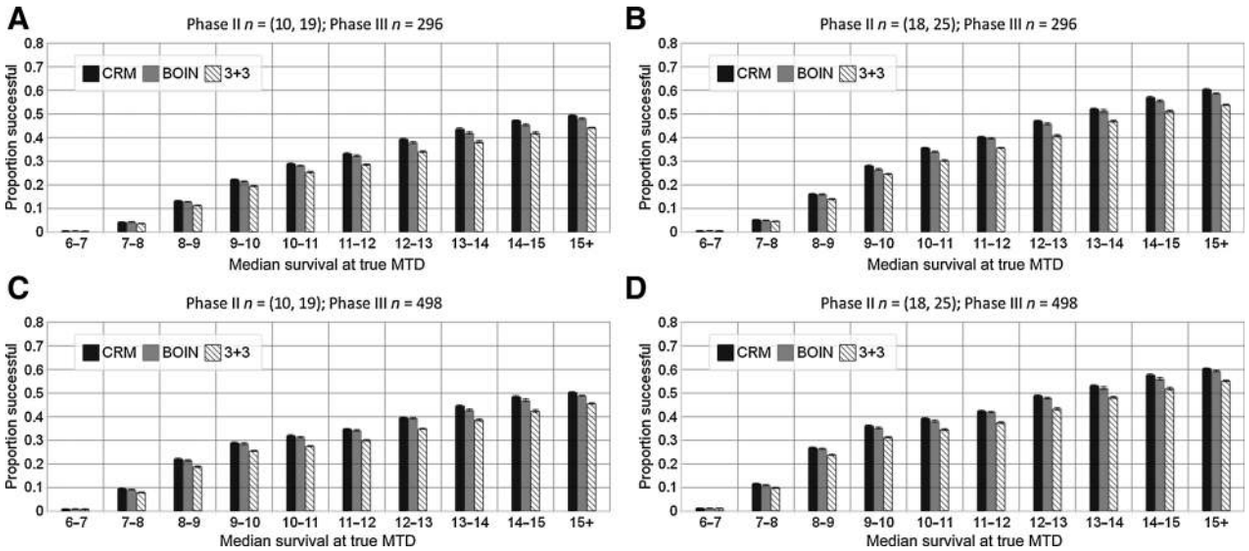

Figure 4.

A–D, Proportion of agents with 2 successful phase III trials among 20,000 drugs (section 3.3) from the population of low-toxicity agents by true median survival at the true MTD.

Discussion

The results presented in this article demonstrate that the use of the CRM, rather than the 3+3 most commonly used, increases the proportion of effective agents shown to be significantly better than the current standard therapy in 2 successive phase III trials. The same is true, though to a lesser extent, for BOIN. In many ways, this analysis underestimates the limitations of the 3+3 relative to the CRM. One of the advantages of the CRM and BOIN is flexibility in setting the target toxicity threshold, for example, increasing or decreasing the threshold depending on the definition of a DLT.

The simulations were carried out on single agents with increasing dose-toxicity, dose-response profiles, and with median survival increasing with dose. Contemporary dose-finding trials often feature combinations of agents, molecular-targeted agents, multiple schedules of agents, and heterogeneous groups of participants or bivariate endpoints involving both response and toxicity. The 3+3 is ill-suited to handle the additional complexities of these studies (24, 25), whereas extensions of the CRM have been proposed to address the challenges of contemporary dose-finding trials, including combinations of agents (26–28), molecular targeted agents (29), participant heterogeneity (30–34), or bivariate outcomes (35, 36). Other model-based extensions have also been proposed to handle the additional complexities of contemporary dose-finding trials (37–47).

In the past, one of the perceived barriers to the use of the CRM was the availability of readily available, easy-to-use statistical software. That perceived barrier has largely been removed with the development of web-based applications to implement and document the operating characteristics of the CRM (48).

We recognize that no simulation study can capture completely the idiosyncrasies of the drug development process. We have proposed a model that we believe captures the fundamental features of drug development, even if not every nuance in the development process. For example, in the simulations, no interim monitoring of the phase III trials is done, whereas in practice, there are usually interim analyses for futility and efficacy. We would not expect this to affect the results presented in this article because the interim analysis boundaries are constructed to allow interim looks at the data without greatly affecting size and power characteristics of the trial. In addition, for the confirmatory phase III trial, we simulated a second trial with the same number of participants as the first. In practice, the size of the second phase III trial would often depend on the observed magnitude of the effect observed in the initial phase III trial.

There have been a number of discussion papers on failure rates of new agents in oncology (49–51). Many of the suggestions revolve around designing larger phase II and phase III trials with multiple endpoints. Gan and colleagues (49) review 235 published phase III cancer trials and find that about 62% of the trials show no significant difference between treatment groups. The authors suggest that much of the reason is overly optimistic estimates of the treatment effect, leading to underpowered trials. We conjecture that part of the reason is that the often-used 3+3 decision strategy tends to be overly conservative and settles on a dose that is not as effective as a higher, but still safe, dose. The smaller effect size at this dose could be part of why new agents are not shown to be superior to a standard therapy. We recognize that this is conjecture on our part, because we do not know what phase I design was used in the 235 studies evaluated by Gan and colleagues, although the majority of phase I trials in general use the 3+3. In contrast, our results indicate that emphasis on more efficient early stage designs, both phase I and phase II, has a much greater effect on the likelihood of success for an agent than the size of the phase III trial.

In an era of greater scientific understanding of the molecular characteristics of cancer and drug targeting, we anticipate that many of the new agents entering the drug development process in the near future will be substantially better than the current

standard. It is an obvious statement that even a well-designed, highly effective agent, with a well understood mechanism of action, cannot benefit participants unless that agent successfully navigates the drug development process. Our results underscore the importance of well-designed early stage designs. Use of better early-phase designs has been advocated by statisticians for many years, but these recommendations have largely gone unheeded. Our models demonstrate that suboptimal dose finding has effects that ripple through the entire development process and suggest that more attention paid to early stage trial design would improve the overall drug development process.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Authors' Contributions

Conception and design: M.R. Conaway, G.R. Petroni
Development of methodology: M.R. Conaway, G.R. Petroni
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M.R. Conaway, G.R. Petroni
Writing, review, and/or revision of the manuscript: M.R. Conaway, G.R. Petroni
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.R. Conaway, G.R. Petroni

Acknowledgments

Research reported in this publication was supported in part by the NCI of the NIH under award number R01CA142859 (M.R. Conaway and G.R. Petroni) and the University of Virginia Cancer Center, University of Virginia Health System P30 CA044579 (M.R. Conaway and G.R. Petroni). The authors thank the referees and the Associate Editor for their insightful comments, which greatly improved the clarity, focus, and accuracy of the revised article. The populations of hypothetical agents described in sections 2.5 and 2.6 are available at http://faculty.virginia.edu/model-based_dose-finding/.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 17, 2018; revised May 7, 2018; accepted October 12, 2018; published first October 16, 2018.

References

1. Storer B. Design and analysis of phase I clinical trials. Biometrics 1989; 45:925–37.
2. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer, Biometrics 1990;46:33–48.
3. Ahn C. An evaluation of phase I cancer clinical trial designs. Stat Med 1998;17:1537–49.
4. Iasonos A, Wilton A, Riedel E, Seshan V, Spriggs D. A comprehensive comparison of the continual reassessment method to the standard 3+3 dose escalation scheme in phase I dose-finding studies. Clin Trials 2008; 5:465–77.
5. Paoletti X, Ezzalfani M, Le Tourneau C. Statistical controversies in clinical research: requiem for the 3+3 design for phase I trials. Ann Oncol 2015;26:1808–12.
6. Cheung Y-K. Dose finding by the continual reassessment method. Boca Ration, FL: Chapman and Hall/CRC Biostatistics Series; 2011.
7. Liu S, Yuan Y. Bayesian optimal interval designs for phase I clinical trials. J R Stat Soc Ser C Appl Stat 2015;32:2505–11.
8. Yuan Y, Hess K, Hilsenbeck S, Gilbert R. Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials. Clin Cancer Res 2016;22:4291–301.
9. Horton B, Wages N, Conaway M. Performance of toxicity probability interval based designs in contrast to the continual reassessment method. Stat Med 2017;36:291–300.
10. Zhou H, Yuan Y, Nie L. Accuracy, safety and reliability of novel phase I trial designs. Clin Cancer Res 2018a;24:4357–64.
11. Zhou H, Murray T, Pan H, Yuan Y. Comparative review of novel model-assisted designs for phase I clinical trials. Stat Med 2018;37:2208–22.
12. Herbst R, Baas P, Kim D, Felip E, Pérez-Gracia J, Han J-Y, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet 2016;387:1540–50.
13. Chatterjee M, Turner D, Felip E, Lena H, Cappuzzo F, Horn L, et al. Systematic evaluation of pembrolizumab dosing in participants with advanced non-small-cell lung cancer. Ann Oncol 2016;27:1291–8.
14. Reck M, Rodríguez-Abreu D, Robinson A, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer. N Engl J Med 2016;375:1823–33.
15. Lin Y, Shih W. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. Biostatistics 2001;2:203–15.
16. Ananthakrishnan R, Green S, Chang M, Doros G, Massaro J, LaValley M. Systematic comparison of the statistical operating characteristics of various phase I oncology designs. Contemp Clin Trials Commun 2017;5:34–48.
17. O'Quigley J, Shen L. Continual reassessment method: a likelihood approach. Biometrics 1996;52:673–84.
18. Lee S, Cheung K-Y. Model calibration in the continual reassessment method. Clin Trials 2009;6:227–38.
19. Agresti A, Coull B. Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 1998;52:119–26.
20. Iasonos A, O'Quigley J. Early phase clinical trials—are dose expansion cohorts needed? Nat Rev 2015;12:626–8.
21. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials 1989;10:1–10.
22. Petroni G, Wages N, Paux G, Dubois F. Implementation of adaptive methods in early-phase clinical trials. Stat Med 2017;36:215–24.
23. DiMasi JA, Reichert JM, Feldman L, Malins A. Clinical approval success rates for investigational cancer drugs. Clin Pharmacol Ther 2013;9: 329–35.
24. O'Quigley J, Conaway M. Continual reassessment and related dose-finding designs. Stat Sci 2010;25:202–16.
25. O'Quigley J, Conaway M. Extended model based designs for more complex dose finding studies. Stat Med 2011;30:2062–9.
26. Wages N, Conaway M, O'Quigley J. Continual reassessment method for partial ordering, Biometrics 2011;67:1555–63.
27. Wages N, Conaway M, O'Quigley J. Dose-finding design for multi-drug combinations. Clin Trials 2011;8:380–9.
28. Hirakawa A, Wages N, Sato H, Matsui S. A comparative study of adaptive dose-finding designs for phase I oncology trials of combination therapies. Stat Med 2015;34:3194–213.
29. Wages NA, Tait C. Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. J Biopharm Stat 2015;25:903–20.
30. O'Quigley J, Shen L, Gamst A. Two sample continual reassessment method. J Biopharm Stat 1999;9:17–44.
31. Yuan Z, Chappell R. Isotonic designs for phase I cancer clinical trials with multiple risk groups. Clin Trials 2004;1:499–508.
32. Conaway MR. A design for phase I trials in completely or partially ordered groups. Stat Med 2017;36:2323–32.
33. O'Quigley J.Phase I and phase I/II dose finding algorithms using continual reassessment method. In: Crowley J, Ankherst D, editors. Handbook of statistics in clinical oncology. 2nd ed. Chapman and Hall/CRC Biostatistics Series; 2006.
34. O'Quigley J, Iasonos A. Bridging solutions in dose-finding problems. J Biopharm Stat 2014;6:185–97.
35. O'Quigley J, Hughes M, Fenton T. Dose-finding design for HIV studies. Biometrics 2011;57:1018–29.

36. Wages NA, Read PW, Petroni GR. A phase I/II adaptive design for heterogeneous groups with application to a stereotactic body radiation therapy trial. Pharm Stat 2015;14:302–10.
37. Conaway M, Dunbar S, Peddada S. Designs for single or multiple agent phase I trials, Biometrics 2004;60:661–9.
38. Thall P, Millikan R, Mueller P, Lee S. Dose-finding with two agents in phase I oncology trials. Biometrics 2004;59:487–96.
39. Yin G, Yuan Y. Bayesian dose finding in oncology for drug combinations by copula regression. J R Stat Soc Ser C Appl Stat 2009;58:211–24.
40. Zang Y, Lee JJ, Yuan Y. Adaptive designs for identifying optimal biological dose for molecularly targeted agents. Clin Trials 2014;11:319–27.
41. Sato H, Hirakawa A, Hamada C. An adaptive dose-finding method using a change-point model for molecularly targeted agents in phase I trials. Stat Med 2016;35:4093–109.
42. Legezda A, Ibrahim J. Heterogeneity in phase I clinical trials: prior elicitation and computation using the continual reassessment method. Stat Med 2001;20:867–82.
43. Babb J, Rogatko A. Participant specific dosing in a cancer phase I clinical trial. Stat Med 2001;20:2079–90.
44. Ivanova A, Wang K. Bivariate isotonic design for dose-finding with ordered groups. Stat Med 2006;25:2018–26.
45. Thall P, Russell K. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. Biometrics 1998;54:251–64.
46. Thall P, Cook J. Dose-finding based on efficacy–toxicity trade-offs. Biometrics 2004;60:684–93.
47. Zhong W, Koopmeiners JS, Carlin BP. A trivariate continual reassessment method for phase I/II trials of toxicity, efficacy, and surrogate efficacy. Stat Med 2012;31:3885–95.
48. Wages N, Petroni G. A web tool for designing and conducting phase I trials using the continual reassessment method. BMC Cancer 2018;18:133.
49. Gan H, You B, Pond G, Chen E. Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. J Natl Cancer Inst 2012;104:590–8.
50. Amiri-Kordestani L, Fojo T. Why do phase III clinical trials in oncology fail so often? J Natl Cancer Inst 2012;104:568–9.
51. Printz C. Failure rate: why many cancer drugs don't receive FDA approval, and what can be done about it. Cancer 2015;121:1529–30.