Impact of Dose Selection Strategies Used in Phase II on the Probability of Success in Phase III

Zoran Antonijevic, José Pinheiro, Parvin Fardipour, and Roger J. Lewis

The purpose of this study was to assess the impact of phase II dose-selection strategies on the likelihood of success of phase III clinical programs, comparing both traditional and adaptive approaches.

We evaluated the impact of the phase II approach to dose selection (including traditional, design-adaptive, and analysis-adaptive approaches), the sample size used in phase II, the number of doses studied in phase II, and the number of doses selected to advance into phase III on the probability of demonstrating efficacy, of demonstrating a lack of toxicity, of phase III trial success, and on the probability of overall success of the combined phase II/phase III programs. The expected net present value was used to quantify the financial implications of different strategies.

We found that adaptive dose allocation approaches (in particular, the Bayesian general adaptive dose allocation method) usually outperformed other fixed dose allocation approaches with respect to both probability of success and dose selection. Design-adaptive approaches were more efficient than analysis-adaptive approaches. The allocation of additional resources into phase II improved the probability of success in phase III and the expected net present value. Bringing two doses forward into phase III testing also increased the probability of success and improved the expected net present value. The overall probability of success in phase III ranged from 35% to 65%, consistent with recent industry experience. This success rate could likely be improved with additional investment in phase II, the use of design-adaptive dose-finding designs when possible, increasing the power of phase III trials, more explicit consideration of toxicity concerns, and better dose selection.

Key Words: Adaptive dose-ranging; Clinical trial simulation; Cost-benefit; Drug development scenarios.

1. Background

The overall success rate of drug development programs has been decreasing, including the success rate of phase III clinical trials. Selection of one or more doses to advance into phase III clinical trials is one of the most challenging decisions during drug development. The increasing attrition rate in phase III is likely due, in part, to faulty dose selection; namely, selecting a dose that is too low to achieve the desired benefit or one that is too potent, resulting in unacceptable toxicity. Dose selection affects more than the probability of regulatory approval. A well-selected dose will have a more desirable risk/benefit profile and thus will result in a greater market value for the product, improved patient care, and greater benefit to society.

In 2005, the Pharmaceutical Innovation Steering Committee of the Pharmaceutical Research and Manu-

Statistics in Biopharmaceutical Research
November 2010, Vol. 2, No. 4
DOI: 10.1198/sbr.2010.08101

facturers of America (PhRMA) formed several working groups to investigate the decreasing success rates of drug development programs. One of these groups, the Adaptive Dose-Ranging Studies (ADRS) working group, was created to analyze existing and to develop new adaptive dose-ranging methods. In November 2007, the *Journal of Biopharmaceutical Statistics* published the group's first product, a comprehensive phase II simulation study comparing performance of adaptive phase II dose-ranging methods, entitled "Innovative Approaches for Designing and Analyzing Dose-Ranging Trials" (Bornkamp et al. 2007). Trial performance was measured by multiple parameters: ability to detect a dose response, identifying a clinically relevant treatment effect, accuracy in selecting a target dose, and estimation of the dose-response curve.

This article builds on this prior work to assess the impact of phase II dose-selection strategies on the success of phase III clinical programs, comparing both traditional and adaptive approaches. Specifically, we consider the probability of success for demonstrating efficacy, the probability of success for demonstrating lack of toxicity, and the overall probability of success in phase III. Phase II approaches are also compared based on the probability of success of regulatory approval after phase III and expected profits as assessed by expected net present value (NPV).

During the design of an adaptive dose-ranging study, multiple choices must be made that significantly impact the probability of success, supply requirements, logistical matters, and resulting costs. We consider a number of these choices below.

1.1 Statistical Methodology

The most accurate statistical methods for selecting the best dose to advance into phase III will yield the highest probability of success in phase III. The choice of methodology will also have financial and logistical implications since methods that allow more flexibility, although potentially more efficient, are also more difficult to implement.

1.2 Number of Doses to be Studied

Traditional approaches to phase II development rely on designs that include up to several doses and a control. In some therapeutic areas, such as oncology, only one dose may be studied. Unfortunately, at the beginning of phase IIb one usually has very limited data on the clinical endpoints, or even biomarkers, as well as very limited information on toxicity. Thus, doses that are selected for phase II evaluation may be too low to achieve maximum efficacy, or may be in a region of the dose-response where the efficacy has reached a plateau and toxicities predominate. Alternatively, the spacing between the doses may be too wide (Grieve and Krams 2005), and a dose with an optimal balance of efficacy and toxicity may lie between two adjacent doses.

On the other extreme, with designs like Bayesian general adaptive dose allocation (GADA), one may be tempted to study a large number of doses (e.g., the ASTIN trial; Grieve and Krams 2005). Accordingly, one should ask if, for any given situation, there is an optimal number of doses to be studied so that, beyond that number, learning about the dose-response is not improved or may even be less efficient. Furthermore, the impact of including many doses on the trial's logistical requirements and the cost of supplies should be considered.

Others have evaluated the impact of the number of doses included in phase II (Bornkamp et al. 2007; Ivanova, Bolognese, and Perevozskaya 2008). Specifically, Bornkamp et al. (2007) considered three different scenarios, namely five equally spaced doses, seven doses, and nine equally spaced doses. In this study we will consider two options: five equally spaced doses and nine equally spaced doses.

1.3 Number of Interim Assessments

Phase II studies often benefit from flexibility in their design. During the process of "searching" for the best dose, it is desirable to be able to discontinue doses with apparently inadequate efficacy or with excess toxicity, or to add doses if the efficacy plateau has not been reached and there is no apparent toxicity even with the highest dose studied. Additionally, some designs like GADA and the D-optimal response-adaptive approach (Bornkamp et al. 2007) allow for adjustment in the allocation ratio based on data observed in the trial itself, so that future patients are preferentially allocated to the most relevant parts of the dose-response curve. The number of interim assessments can vary widely; the key question is whether there is an optimal number of interim assessments, either from the point of view of statistical efficiency or to balance statistical efficiency with logistical difficulty. In the current study, however, for reasons of practicality, we considered the frequency of interim assessments to be out of our scope.

1.4 Optimal Size of Phase II Relative to Phase III

Calculating the sample size required for an adaptive phase IIb trial can be quite complex. The primary objective of a phase IIb trial is to select the "best" dose(s) for use in a confirmatory trial, where the definition of the best dose is based on an optimal balance of efficacy and toxicity. In addition, during phase IIb, one also hopes to obtain a good estimate of the efficacy of the selected dose(s), to be used in calculating the sample size required for subsequent phase III trial(s). Trade-offs associated with a smaller sample-size investment in phase IIb include: (1) more doses may have to be studied in the confirmatory phase due to insufficient knowledge of the dose response, increasing phase III cost and complexity; (2) a larger sample size may be necessary in the confirmatory phase due to uncertainty regarding the treatment effect, to avoid having the confirmatory trial fail for lack of demonstrated efficacy (Type II error); and (3) a greater probability of having to discontinue a dose in phase III due to previously unappreciated toxicity. On the other hand, a larger phase IIb study may both increase costs and delay the initiation of phase III. A long delay can also reduce the expected net present value (NPV) of the product.

Bornkamp et al. (2007) addressed the impact of two sample sizes, 150 and 250 subjects, on multiple phase IIb outcomes. We extend that prior work by addressing the following questions: (1) what is the impact of the phase II sample size on the probability of success in the phase III; and (2) what is the impact of the phase II sample size on the product's expected net present value?

1.5 Dose-Selection Criteria for Phase III

The selection criteria for selecting the dose(s) to be carried forward into phase III may be based solely on efficacy or on a balance of efficacy and toxicity. The reliability of any criteria is limited by the amount of information available in phase II. This limitation is particularly important in assessing toxicity, given that toxicity information is usually accrued at a slower rate than information on efficacy. If the dose-selection criteria are based on efficacy only, then one may choose a dose with a minimum clinically significant difference or a dose beyond which there is no meaningful improvement in efficacy. Toxicity considerations can then be incorporated into the decision criteria qualitatively, perhaps by using an independent data monitoring committee (IDMC) to provide a dose selection recommendation to the sponsor. Another method for incorporating toxicity considerations into dose selection criteria is to create a utility function which combines the positive benefits from efficacy and the negative effects of toxicity (Berry et al. 2001; Dragalin and Fedorov 2006). Finally, toxicity considerations can be incorporated into decision criteria implicitly, by selecting a dose with a certain percent of the maximum efficacy, for example, a dose delivering 95% of the maximum efficacy, denoted the ED95 (Grieve and Krams 2005).

1.6 Number of Doses to Take into Phase III

While advancing a single dose into phase III is appealing and allows a simpler trial design, there are reasons to consider advancing more than one dose into the confirmatory stage (Hemmings 2007). Sometimes the difference in efficacy between two doses in phase II is too small to allow a reliable choice, or an unexplained or implausible inverse dose relationship might have been observed. More often, the toxicity data collected during phase II may be inconclusive. It may be unclear after phase II if the efficacy of the lower doses is sufficiently good to warrant an approval while, simultaneously, there may remain some concerns associated with the toxicity of the higher doses. For many indications, the increase in the expected revenue from an increased probability of product approval may be greater than the additional costs incurred by considering another dose in phase III. In order to address this issue, we have conducted a formal cost analysis.

2. Methods

2.1 Objectives

The purpose of this study is to assess the impact of phase II design characteristics on the probability of success in phase III (defined as the probability of two successful, pivotal confirmatory trials, as usually required for regulatory approval) and on the expected net present value of the product. The impacts of the following phase II characteristics were studied: (1) the statistical approach to dose selection; (2) the sample size used in phase II; (3) the number of doses studied in phase II; and (4) the number of doses selected to advance into phase III.

2.2 Scenarios

To include a broad range of scenarios in the simulations, we considered all seven statistical approaches to phase II dose selection described by Bornkamp et al. (2007), four efficacy/toxicity dose-response profiles (see Figure 1), two phase II sample sizes (total of 150 and 250 patients), two numbers of doses to be included in phase II (5 and 9), and either one or two doses selected in phase II to advance into phase III.

2.3 Primary Endpoint

As in the study by Bornkamp et al. (2007), the example indication used in this study was neuropathic pain and the primary endpoint was a change in pain from baseline to Week 6, as measured by a Visual Analog Scale (VAS).

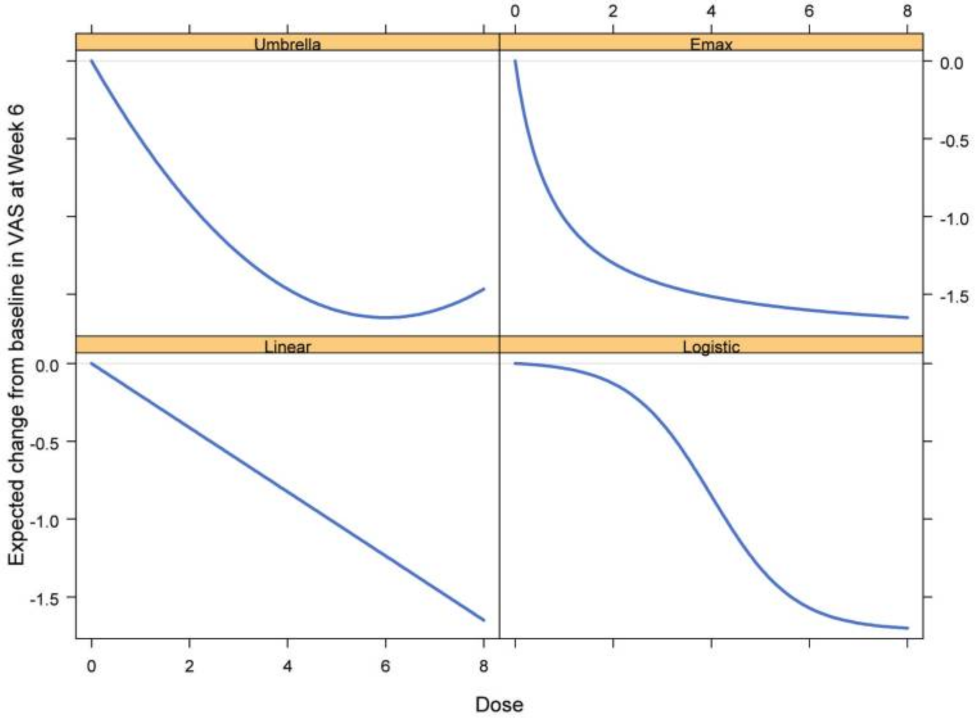

Figure 1. Assumed efficacy-dose response profiles. This figure shows the four dose-response relationships assumed in the simulations, denoted umbrella (quadratic), $E_{\max}$, linear, and logistic. These represent a subset of the dose-response relationships considered by Bornkamp et al. (2007).

The VAS takes values between 0 (no pain) and 10 (highest pain) on a continuous scale (Gallagher, Liebman, and Bijur 2001).

2.4 Phase II Design

2.4.1 Statistical Approaches

In this study we focused on the statistical approaches to dose selection that were described and evaluated by Bornkamp et al. (2007). These approaches can be classified into three broad groups:

1. *Traditional*: The traditional approach relies on analysis of variance (ANOVA) to compare each dose's response to control, using Dunnett's adjustment (Dunnett 1955) to control for multiplicity. The data analysis is completely prespecified and no adaptation is used in either the design or the analysis.

2. *Adaptive Design*: Approaches that allow study design parameters to be changed during trial conduct based on the data collected within the trial. Two design-adaptive approaches were studied: (1) Bayesian general adaptive dose allocation (GADA); and (2) the D-optimal response-adaptive approach (D-Opt). The GADA method uses Bayesian dose-response modeling to borrow information across nearby doses and longitudinal modeling to use all available information from subjects with incomplete data. Dose allocation is based on minimizing the variance of the parameter of interest (e.g., the response at the target dose). In the D-Opt method, allocation proportions are adapted in a group sequential manner to maximize the expected information over the dose-response curve, according to the D-optimality criterion.

3. *Adaptive Analysis*: Approaches in which the best method for analysis is driven by the data collected during the trial, but there are no response-adaptive changes in dose allocation during the trial. The following adaptive analysis methods were studied: (1) a combination of modeling and multiple comparisons procedures proposed by Bretz, Pinheiro, and Branson (2005), denoted MCP-Mod; (2) a multiple trend test (MTT) approach based on selecting three curves (upper, lower, and middle) from a class of sigmoid $E_{\max}$ models to minimize the power of the associated triple trend test; (3) Bayesian model averaging (BMA) in which a set of relatively simple dose response models, and priors that include both model weights and the model parameters them-

selves, are updated using standard Bayesian inference to obtain posterior estimates for the dose-response; and (4) a nonparametric linear regression approach (LOCFIT) that relies on model-free, nonparametric regression, with a local quadratic regression technique used for the dose-selection step.

More detailed descriptions of each of these methods can be found in Bornkamp et al. (2007).

2.4.2 Dose-Response Profiles

Four assumed dose-response relationships for efficacy were simulated. These are denoted logistic, linear, umbrella, and $E_{\max}$. The four assumed dose-response curves were

1. *Logistic*: $\Delta\text{VAS} = 0.015 - 1.73/(1 + \exp(1.2 * (4 - \text{dose}))) + \varepsilon$
2. *Linear*: $\Delta\text{VAS} = -(1.65/8) * \text{dose} + \varepsilon$
3. *Umbrella*: $\Delta\text{VAS} = -(1.65/3) * \text{dose} + (1.65/36) * (\text{dose2}) + \varepsilon$
4. $E_{\max}$: $\Delta\text{VAS} = -1.81 * \text{dose}/(0.79 + \text{dose}) + \varepsilon$

where ΔVAS denotes the change in pain, as assessed by the visual analog scale (VAS) from baseline to six weeks of treatment and ε represents the random error. These dose-response profiles are a subset of those considered by Bornkamp et al. (2007) and are shown graphically in Figure 1.

If toxicity were not a factor then, for any monotonic efficacy response, a higher dose would always be more successful without any upper limit on the dose. This is not the case in practice and, accordingly, we have imposed a toxicity penalty on increasing doses selected for inclusion in phase III. The toxicity penalty function reflects the probability of a treatment-limiting toxicity being detected for a patient in a phase III trial. Four different toxicity penalty functions were applied, with each theoretical efficacy profile having a corresponding toxicity penalty function. It was also assumed that the dose-response for the probability of detecting excess toxicity in phase III would become steeper at higher doses. Specifically, we assumed:

- For placebo, the assumed probability of a treatment-limiting toxicity occurring for a patient in phase III was 5%.
- The probability of a treatment-limiting toxicity occurring for a patient in phase III for the target dose was 6%. The target dose for each assumed efficacy dose-response profile was the dose that yielded a separation from placebo of -1.3 units in the VAS. Since the dose in a dose-response model is defined on the continuous scale, the target dose was rounded to the nearest integer.
- The toxicity penalty function (i.e., the probability of treatment-limiting toxicity occurring for a patient in phase III) was assumed to be linear between placebo and the target dose.
- At the dose just above the target dose, the assumed probability of treatment-limiting toxicity was 7.5%, unless this dose is also the maximum dose studied (dose 8).
- The probability of a treatment-limiting toxicity increased to 12% at the maximum dose (dose 8).
- The toxicity penalty function was exponential between the dose just above the target dose and the maximum dose.

The resulting dose-response relationships, for the probability of a treatment-limiting toxicity occurring for a patient in phase III, are shown in Figure 2, with each toxicity dose-response profile corresponding to an efficacy dose-response profile in Figure 1. The toxicity dose-response profiles were constructed on an ad-hoc basis to more severely penalize the selection of doses above the upper limit of the target dose intervals defined in Bornkamp et al. (2007). The goal was to favor the selection of doses within the target dose interval corresponding to underlying dose-response efficacy model.

2.4.3 Phase II Sample Sizes

We used simulated sample sizes for the phase II trials of 150 and 250 subjects, to mirror the prior work of Bornkamp et al. (2007).

2.4.4 Number of Doses Studied in Phase II

Bornkamp et al. (2007) studied the performance of phase II designs with five equally spaced, seven unequally spaced, or nine equally spaced doses. For the work here, for reasons of practicality, we limited our simulations to either five or nine equally spaced doses. Specifically, the five doses were 0, 2, 4, 6, and 8 with "0" denoting placebo and the other values denoting the dose in an appropriate unit (e.g., milligrams). The nine doses were the integers from 0 to 8, again with "0" denoting placebo and the others the dose in an appropriate unit.

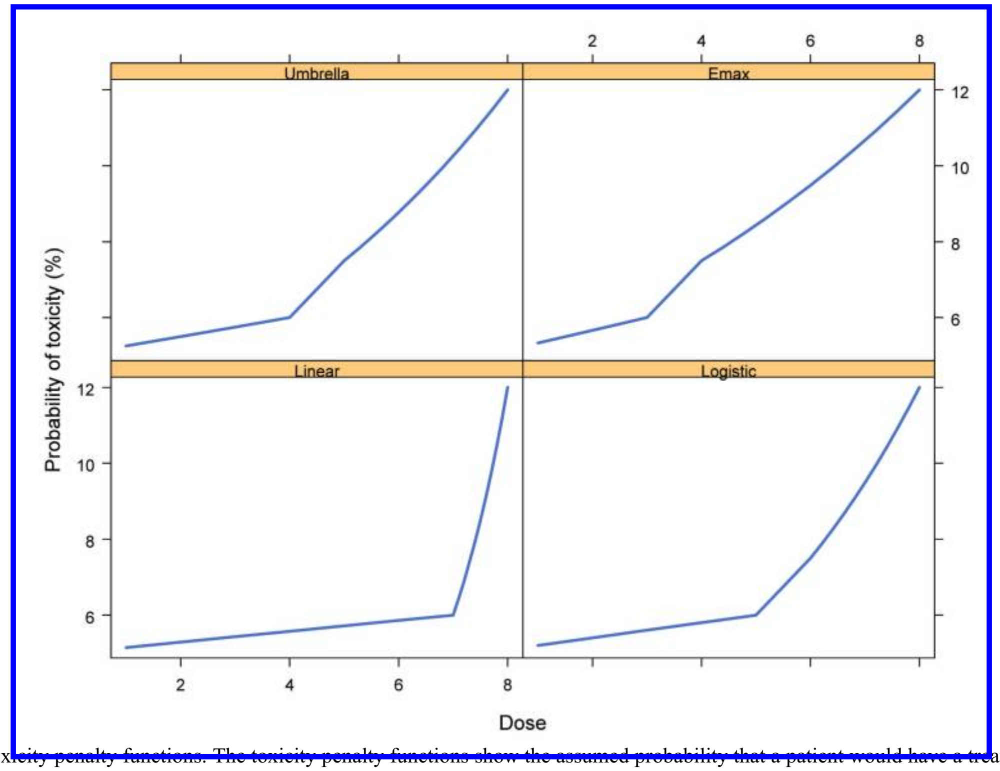

Figure 2. Toxicity penalty functions. The toxicity penalty functions show the assumed probability that a patient would have a treatment-limiting toxicity during a phase III trial, as a function of the dose selected for study in phase III and the underlying assumed efficacy dose-response relationship.

2.4.5 Dose Selection Strategy

For dose selection, results from Bornkamp et al. (2007) were used. These results included the estimated dose-response for each combination of statistical dose-selection method, assumed dose-response, phase II sample size, and the number of doses studied. In the original work, the simulations had been repeated a minimum of 5,000 times for each scenario.

As a first step, we estimated the change from baseline (difference between the response to any given dose and placebo) in each simulated phase II trial. We then selected the dose according to the rule associated with the phase II statistical approach being evaluated. The dose selection rules are detailed in Bornkamp et al. (2007). The dose selection step occurred separately for each individual simulated phase II trial and since each simulated trial was based on a different realization of simulated trial data, different simulated trials might yield different selected doses. Thus, even for a particular combination of statistical dose-finding approach, phase II trial size, assumed dose-response, etc., and with a single-dose phase III design, we considered quantitatively the fact that different doses might be brought forward into phase III. Since each pair of phase III trials that constituted a phase III program were always based on the same simulated phase II trial, they always included the same dose for evaluation.

To evaluate phase III designs with two active doses, the first dose was selected as described above. For the second dose, we considered the dose immediately above and the one immediately below the first selected dose, selecting the one whose response was closer to the target efficacy, the minimum clinically meaningful difference (MCMD). Since each pair of phase III trials that constituted a phase III program were always based on the same simulated phase II trial, they always included the same two doses for evaluation.

2.5 Phase III Design

Our modeling of the phase III program assumed that two pivotal confirmatory studies were necessary for a successful submission. Thus, phase III success was defined using a two-sided $\alpha = 0.05$ on both trials, this being a standard regulatory requirement. We then considered two phase III designs, with either one or two active dose arms with the doses selected as described above. Dunnett's procedure (Dunnett 1955) was applied to control for multiplicity for studies with two active dose arms.

2.5.1 Sample Size for the Phase III Study

For phase III trials with one active dose arm, the sample size was calculated to be 86 patients per arm, assuming the MCMD of -1.3 with a standard deviation (σ) of 2.6, a two-sided α=0.05, and a power of 90%. These assumptions were consistent with those used by Bornkamp et al. (2007) except that the variance (σ^2) was inflated by 50%, under the assumption that the phase III populations would be more diverse than those enrolled in phase II.

For phase III trials with two active dose arms, the sample size was calculated to be 99 patients per arm. The assumed MCMD, power, and standard deviation were the same as for the phase III trial with one dose. To control the Type I error rate, Dunnett's adjustment was applied (Dunnett 1955), resulting in a two-sided $\alpha = 0.027$. It was also assumed that one of the doses was going to fail, so that the power determination was based on a single arm. This results in a conservative estimate for the required sample size, equivalent to powering the trial for one active arm using a two-sided $\alpha = 0.027$, but including an additional dose.

2.5.2 Probability of Success

For any given selected dose and corresponding dose response model, we calculated the theoretical "true" treatment effect. To evaluate the phase III designs with a single active dose, the assumed theoretical response for each selected dose was used to determine the power of the phase III trial that would follow, based on the sample size, α level, and this "true" effect size. This calculated power represents the probability of success in demonstrating efficacy.

The rule for determining unacceptable toxicity was based on the difference in the number of patients experiencing treatment-limiting toxicities between an active dose and placebo arms. For the single active dose case ($N = 86$/arm) the critical value for the rule was 4 (i.e., if the difference in number of patients experiencing treatment-limiting toxicity was 4 or more, the dose would fail due to unacceptable toxicity). The critical value for the two active dose case ($N = 99$/arm) was 3. Those critical values were chosen to ensure a small chance ($<5\%$) of failing a dose when its treatment-limiting toxicity probability was the same as placebo, but sufficiently high probability ($> 70\%$) of rejecting doses with treatment-limiting toxicity probabilities of 12% or higher. The probability of success in demonstrating lack of toxicity was then derived for each selected dose(s) under a given toxicity dose-response profile, with respect to the appropriate rule. The calculation of the probability of success was then done assuming that successes in demonstrating efficacy and a lack of toxicity were independent. While this is a strong and potentially unrealistic assumption, it allows a qualitative assessment of the impact of the various factors considered in simulation scenarios. Further evaluations incorporating stochastic dependence between efficacy and toxicity will be a topic of future research.

For phase III trials with two active dose arms, we declared success if at least one of the doses had significant efficacy and lacked unacceptable dose-limiting toxicity. The multiplicity correction to the Type I error rate, using Dunnett's procedure, was applied for efficacy. If one dose was successful for efficacy only, while only the other dose lacked unacceptable dose-limiting toxicity, then the study was declared a failure. Likewise, the phase III program was only declared a success if one of the selected doses demonstrated both efficacy and a lack of unacceptable toxicity in both studies (i.e., the same dose in both studies).

2.6 Comparisons Based on Financial Measurements

The goal of a drug development program is to bring clinically meaningful enhancements to the armamentarium of treatment options at the earliest possible time. The best design for a single trial is one which provides scientific integrity and validity, and delivers the highest information value per resource unit invested. In other words, we want to make the most of the information provided by each subject participating in the trial. To address these goals, we used simulated scenarios to compare each approach's operating characteristics. In addition, however, one must consider financial implications and, rather than looking at the cost of any one trial in isolation, we must consider the overall objectives of the clinical development plan and make design choices with the end in mind. Some questions to be addressed are:

- Does additional investment in phase II pay off sufficiently, in terms of the improved probability of success and consequent improvement in the expected net present value that it is worth investing more in phase II?
- Similarly, does it pay to invest more in phase III by including more than one drug dose in the confirmatory studies?
- What is the optimal resource allocation to the phase II program relative to the phase III program?

The financial analysis should consider three aspects, namely: (1) the information value of a design—the level of certainty with which the research question is answered; (2) the direct cost of a design—the cost to conduct the trial, including investigator cost, drug supply, and management costs; and (3) the earliest time point at which we acquire sufficient information to make the

Table 1a. Phase II drug development program structures and costs used in determination of net present value

	Traditional phase II		Adaptive phase II	
Criteria	250 Subjects	150 Subjects	250 Subjects	150 Subjects
Enrollment period	12 months	12 months	12 months	12 months
Total trial duration	18 months	18 months	18 months	18 months
No. of sites	50	50	50	50
No. of pages per CRF	260	250	260	250
No. of trials	1	1	1	1
No. of subjects per trial	250	150	250	150
Cost (USD) per each trial	$15,178,016	$12,416,266	$15,293,254	$12,493,409

Table 1b. Phase III drug development program structures and costs used in determination of net present value.

	Phase III		
Criteria	1 Active Dose	2 Active Doses: Normal	2 Active Doses: Fast
Enrollment period	12 months	24 months	12 months
Total trial duration	18 months	30 months	18 months
No. of sites	50	50	100
No. of pages per CRF	260	270	270
No. of active dose arms	1	2	2
No. of trials	2	2	2
No. of subjects per trial	172	297	297
Cost (USD) per each trial	$13,069,977	$17,095,465	$19,019,010

correct decision as to whether to move forward with a regulatory submission.

In order to compare studies based on the expected net present value, the first step was to calculate costs for each individual drug development program, accounting for all direct and indirect costs. Since the medical indication of neuropathic pain was used in our prior examples, cost parameters appropriate for conducting studies in this indication were used. The trial structure and cost assumptions used are presented in Table 1. The estimated costs contain the operating expenditures and resources that would be assigned to each trial. It was assumed that both phase III trials would be conducted in parallel and that the phase III program would not start until phase II was completed.

In developing the estimated costs shown in Table 1, assumptions related to the number of sites, the trial enrolment period, and the number of case record form (CRF) pages were based on trials of a treatment for neuropathic pain that was familiar to one of the authors but proprietary. For resource estimates, a work-load forecast was developed to address the required monthly resources from various functional departments that would be involved in these trials. The estimated resources were calculated to address the activities involved during the set-up period, the active period, and afterwards. Lastly, the business model reflected the incurred costs from phase II onwards and incorporated the period of exclusivity for the compound from the time of registration.

The expected exclusivity period was calculated under the assumption that the drug was novel and was not biologic, and thus would have a 20-year patent life. It was also assumed that it took 4 years in early development, 6 years between the start of the patent life and the beginning of the phase II program, and that registration required one year. We used a discount rate of 1.1/year for the net present value and assumed a tax rate of 37.5%.

We assumed the use of a phase II design with five doses. We considered scenarios in which a larger, two-dose phase III trial would take longer to complete (30 months versus 18 months) and scenarios in which both trials could be completed in 18 months (Tables 1b and 2). The last scenario was included so that the impact of the exclusivity period on the expected net present value could be better assessed.

Finally, the expected revenues were included in the calculation of expected net present value, reflecting the likely market size for our example indication. We assumed the revenue stream would increase linearly to $500 million over the first 5 years, remain constant over the patent life, and yield a decreasing income of $20, $10, and $5 million for the three years after the expira-

Table 2. Estimated years of patent life remaining after approval

Number of Patients in Phase II	Revenue (years)		
	Phase III with 1 active dose	Phase III with 2 active doses: Normal enrollment	Phase III with 2 active doses: Fast enrollment
150	10.5	9.5	10.5
250	10.3	9.3	10.3

tion of the patent. We assumed zero income 3 years after the patent expired. The expected net present values provided are presented as mean values and standard deviation, since each simulation run is associated with a single expected net present value.

3. Results

3.1 Probability of Success in Demonstrating Efficacy

The probability of success in demonstrating efficacy in a phase III trial, as a function of the underlying dose-response, the statistical approach, the sample size in phase II, and the number of doses carried forward into phase III (one or two), is shown in Figure 3. For almost all dose-response profiles, adaptive design approaches in general, and GADA in particular, demonstrated the highest probability of success for efficacy. This was true regardless of the phase II sample size, number of doses studied in phase II, or the number of doses selected to be advanced into phase III. For the $E_{\max}$ dose-response profile and the case in which two doses were carried forward to phase III, however, all methods demonstrated a very strong probability of success. This likely occurred

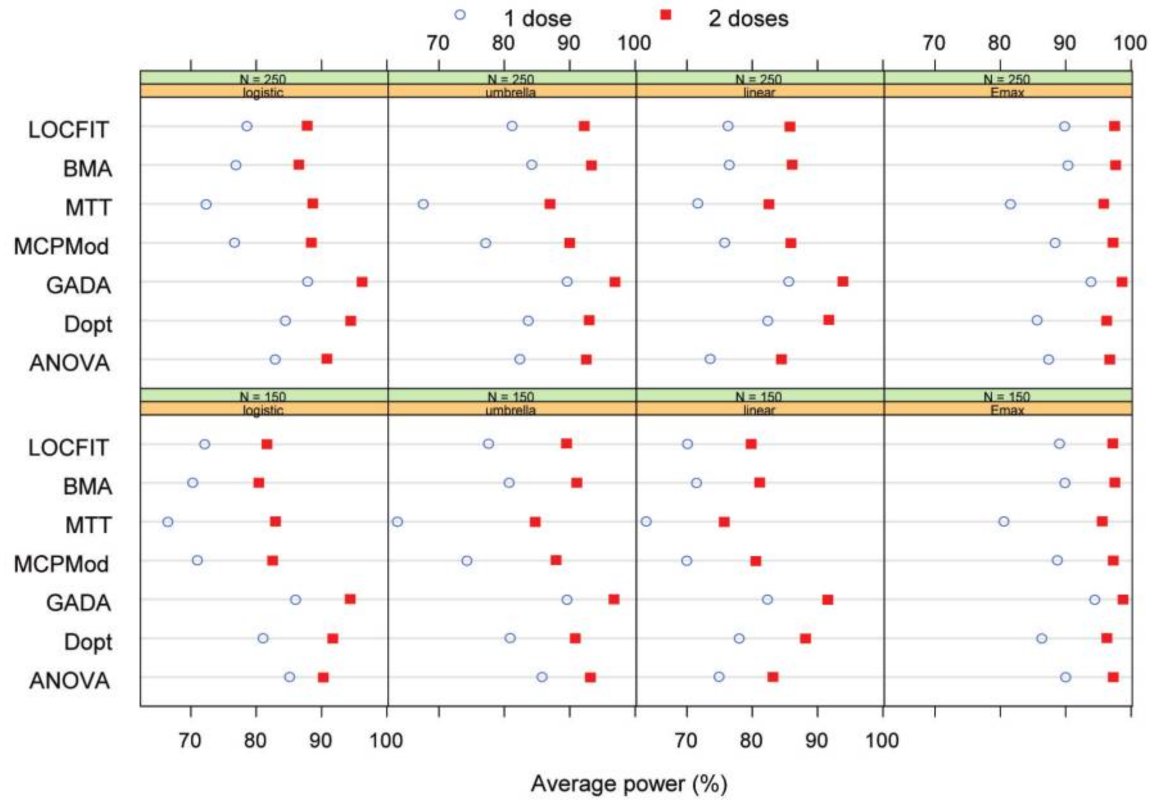

Figure 3. Probability of success in demonstrating efficacy in a phase III trial. This figure shows the probability of success in demonstrating efficacy in a single phase III trial, denoted "average power," as a function of the assumed underlying dose-response profile, the statistical approach used in phase II, the sample size in phase II, and the number of doses carried forward into phase III (one or two).

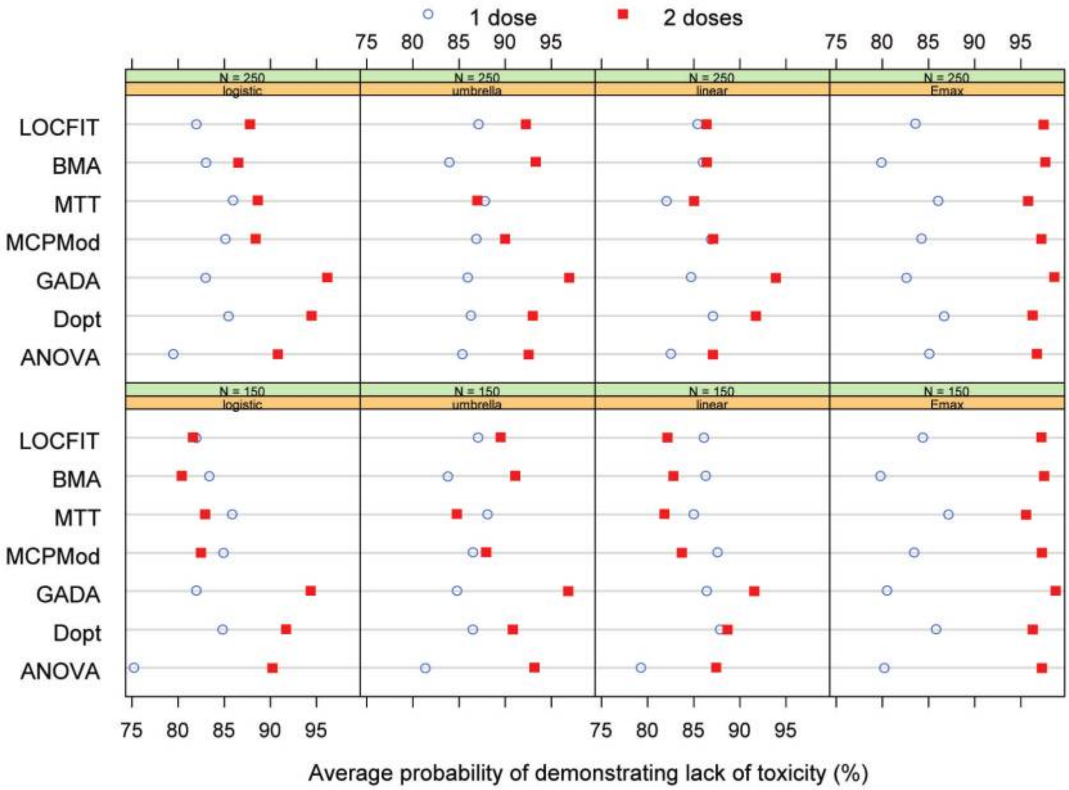

Figure 4. The probability of demonstrating an acceptable toxicity profile in a phase III trial. This figure shows the probability of demonstrating an acceptable toxicity profile in a phase III trial, as a function of the assumed underlying dose-response profile, the statistical approach used in phase II, the sample size in phase II, and the number of doses carried forward into phase III.

because this profile implied a very steep improvement in efficacy with dose, with meaningful efficacy starting as low as a dose of 2. Therefore, when two doses are selected for phase III evaluation, it is very unlikely for both to fail on efficacy with this dose-response profile. Phase II designs with 250 patients show slight, but consistent improvement over designs with 150 patients.

Selecting two active doses to advance into phase III, rather than one, resulted in a consistently higher probability of success for demonstrating efficacy. This may not be surprising, given that we defined success to include the possibility that only one of the two doses demonstrated statistically significant efficacy, and simultaneously used a conservative approach to sample size calculation that incorporated the possibility that one dose would fail.

3.2 Probability of Demonstrating a Lack of Toxicity

Figure 4 shows the probability of demonstrating an acceptable safety profile in a phase III trial, as a function of the underlying dose-response for efficacy (and therefore the dose-response for toxicity), the statistical approach, the sample size in phase II, and the number of doses carried forward into phase III. It should first be noted that toxicity was not part of the dose selection criteria, and therefore any comparison of the methods based on the toxicity of selected doses alone is unlikely to yield useful insights. The probability of avoiding treatment-limiting toxicity is just a reflection of how "high up the dose response curve" various approaches will progress in their search to find a dose with optimal efficacy.

The most striking characteristic of the results for the probability of demonstrating a lack of treatment-limiting toxicity is the gain in the probability of success obtained by selecting two doses to advance to phase III. This gain can be attributed to the "distribution of risk" associated with defining a trial as successful even if only one of the two selected doses demonstrates a sufficiently safe profile. The largest difference in the probability of demonstrating a lack of toxicity between designs with one or two doses carried forward to phase III is for the $E_{\max}$ dose-response profile. Similar to our observations for the probability of success for efficacy, this can be attributed to the large stepwise increase in the risk of treatment lim-

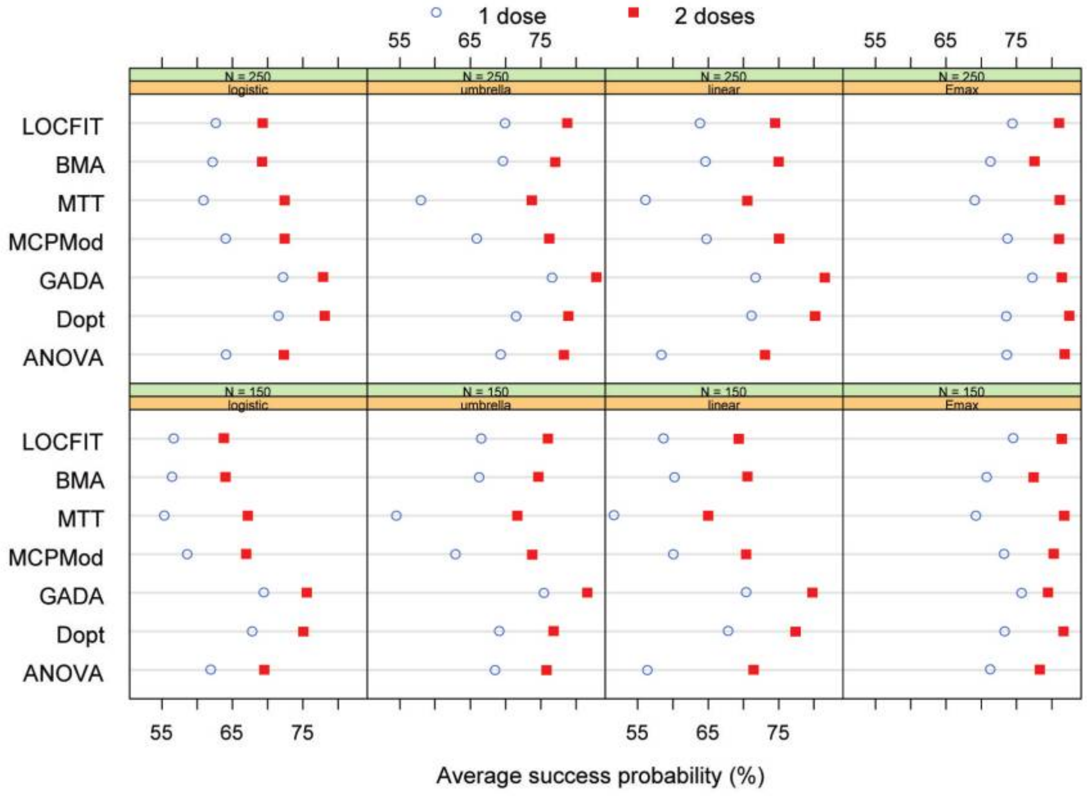

Figure 5. The overall probability of success for an individual phase III study. This figure illustrates the overall probability of success for an individual phase III study, meaning success in demonstrating both efficacy and a lack of treatment-limiting toxicity, as a function of the underlying dose-response profile, the statistical approach used in phase II, the sample size in phase II, and the number of doses carried forward into phase III.

iting toxicity, from one dose to the next, for this dose-response profile compared to the others.

3.3 Overall Probability of Success for a Single Phase III Study

Figure 5 illustrates the overall probability of success for an individual phase III study, meaning success in demonstrating both efficacy and a lack of treatment-limiting toxicity, as a function of the underlying dose-response profile, the statistical approach used in phase II, the sample size in phase II, and the number of doses carried forward into phase III. Adaptive designs in general, and GADA in particular, consistently outperformed other phase II designs and methods. The only exception to this pattern occurred in some cases with the $E_{\max}$ dose-response profile. As we observed with the probability of success for efficacy alone, phase II designs with 250 subjects show consistently better performance than designs with 150 patients.

We observed that phase III designs with two active dose arms demonstrated better probability of success for both efficacy and lack of toxicity when considered separately. Thus it is not surprising, as shown in Figure 5, that bringing two doses forward into phase III, rather than one, increases the probability of demonstrating both efficacy and avoiding toxicity with at least one of the doses. The only way this might not occur would be if there was poor matching of safe and efficacious doses, that is, if two-dose phase III trials tended to find one dose was efficacious but not safe, while the other was safe but not efficacious. For the scenarios examined this is not the case, and bringing two doses forward into phase III improved the overall probability of success. One should note, however, that a total sample size for one of the phase III trials with active two doses is 297, while the sample size for a phase III design with one active dose is 172 (Table 1b). The important question, then, is whether the additional investment in larger phase III trials, as well as in the larger phase II trial, pays off in terms of expected revenues. This issue is addressed below.

3.4 Probability of Success for the Phase III Development Program

For a phase III program to be successful, individual trial successes must be achieved in both phase III trials. The probabilities of success for the phase III program are

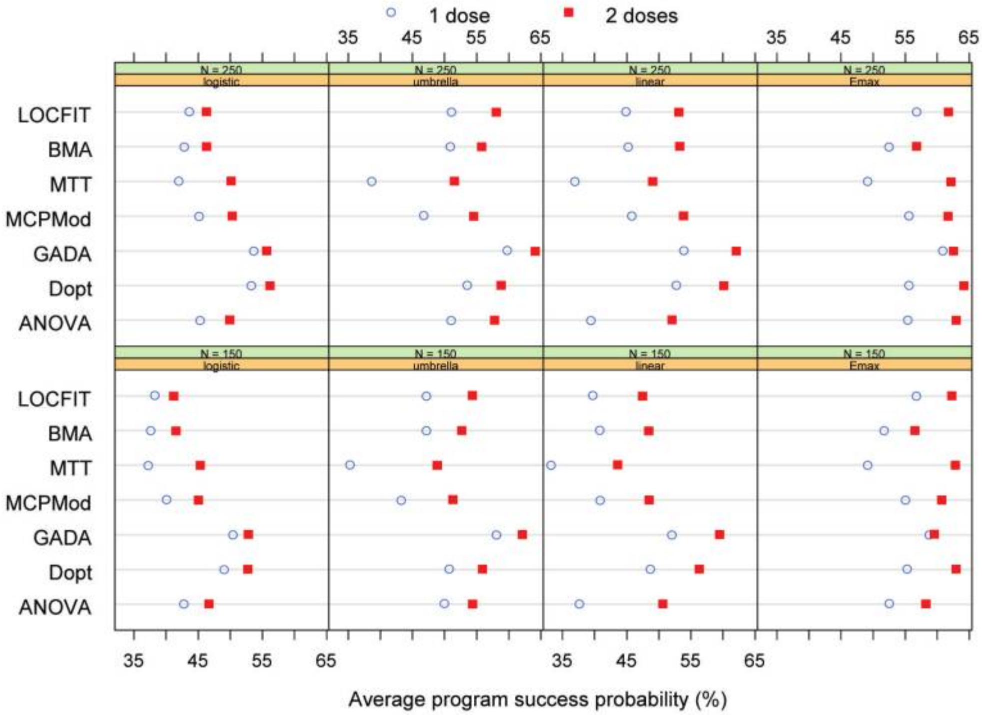

Figure 6. Probability of success for the entire drug development program. This figure demonstrates the probability of success for the entire drug development program, consisting of one phase II and two phase III trials, as a function of the underlying dose-response profile, the statistical approach used in phase II, the sample size in phase II, and the number of doses carried forward into phase III.

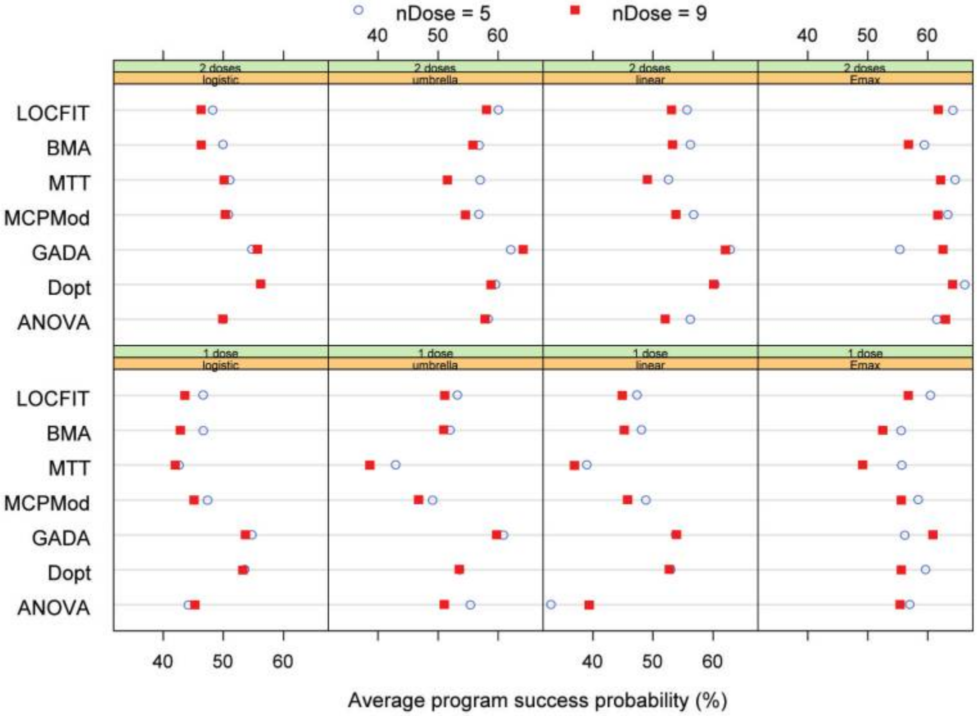

Figure 7. Number of phase II doses considered and the probability of success for the entire drug development program. This figure demonstrates the probability of success for the entire drug development program, consisting of one phase II and two phase III trials, as a function of the underlying dose-response profile, the statistical approach used in phase II, the number of doses (5 or 9) considered in phase II, and the number of doses carried forward into phase III. The number of subjects used in phase II was held constant at 250.

shown in Figure 6. The probabilities of success for the whole program are consistent with those for each phase III study, except that the probabilities are scaled down, reflecting the need for both phase III trials to be successful.

Up to this point we have not explored the effect of changing the number of doses considered in phase II on the probability of success in phase III and the prior results assumed that nine doses were evaluated in phase II. Figure 7 shows the probability of success for the phase III program, consisting of two phase III trials, as a function of the underlying dose-response profile, the statistical approach used in phase II, the number of doses (five or nine) considered in phase II, and the number of doses carried forward into phase III. The number of subjects used in phase II was held constant for these calculations.

In most cases, phase II designs with five treatment arms performed better than those with nine treatment arms. This was particularly the case for designs that do not allow changes in design parameters based on results from interim analyses. It appears that, for designs that do not allow for dropping doses or adapting the allocation ratio, the benefit of studying more doses may be limited and allocating more patients to individual doses may be more efficient. With added flexibility in the design, however, spreading the sample size seems to be more efficient. For the D-Opt method, the performance of designs with five or nine doses is about equal while, for the more flexible GADA design, the performance of designs with nine doses is frequently better than that of the design with five doses.

3.5 Accuracy of Dose Selection

As noted above, the calculated probability of success for our simulated phase III programs were quite low, ranging from 35% to 65%. While a proportion of the phase III failures can be attributed to treatment-limiting toxicity, we wanted to further assess the performance of our dose selection criteria to determine whether poor accuracy or reliability in dose selection was contributing to the phase III failure rate. Our dose selection criteria were straightforward—we chose the dose that was the closest to the MCMD. While the minimum efficacious dose is of interest, the ultimate goal is to select the dose that optimizes the risk/benefit profile. Thus our dose selection criteria could perform poorly because they were neither structured to explicitly choose the dose with the best efficacy nor to explicitly incorporate considerations of toxicity.

Figures 8–11 illustrate the accuracy of the dose selection criteria for the logistic, linear, umbrella, and $E_{\max}$ dose-response models, respectively. In each figure, the dose with the highest overall probability of success is identified, along with the distribution of doses selected by the individual simulated phase II trials. In general, the dose-selection criteria failed to select the dose with the highest probability of success, tending to select a dose lower in strength. The logistic dose-response model was an exception, however, and was associated with more accurate dose selection.

3.6 Expected Net Present Value

The years of patent life remaining after approval, which we term the period of exclusivity, is a key determinant of expected net present value which, of course, is shortened with increasing duration of phase II or phase III evaluation. The assumed enrollment times for the phase II and phase III trial configurations that we have considered are given in Table 1. The resulting expected periods of exclusivity are presented in Table 2, demonstrating that a larger phase II program results in a slightly shorter exclusivity period, while the consideration of a second dose in phase III shortens the period of exclusivity by almost 10%, unless the enrollment can be expedited.

The mean expected net present value associated with each strategy is shown in Figure 12 and the associated standard deviations are shown in Figure 13. An increase in sample size from 150 to 250 in phase II resulted in an increase in the expected net present value for both adaptive and traditional designs. Similarly, considering a second dose in phase III improved the expected net present value after almost all designs, especially if the larger trial could be completed quickly. The improvement of the design with two doses over the design with one dose was the smallest for GADA. This may suggest that use of more efficient phase II adaptive designs would require less investment in phase III. The GADA and D-Opt methods resulted in the largest net present value, and often the lowest variability, for the logistic, quadratic/umbrella, and linear dose-response curves, but not for $E_{\max}$. The GADA method with one dose in phase III often outperformed all nonadaptive phase II methods with either one or two doses used in phase III. The phase III trial with two doses and fast enrollment results in steady improvement in the expected net present value over the trial with normal enrollment despite higher costs associated with implementing such a trial.

4. Discussion

4.1 Factors Influencing the Probability of Phase III Success

One striking finding from our simulations is the disappointingly low probability of success for phase III drug development programs, even under the assumption that the compound being evaluated has clinically important efficacy. Although we powered our studies at 90%, the probability of phase III program success as a whole

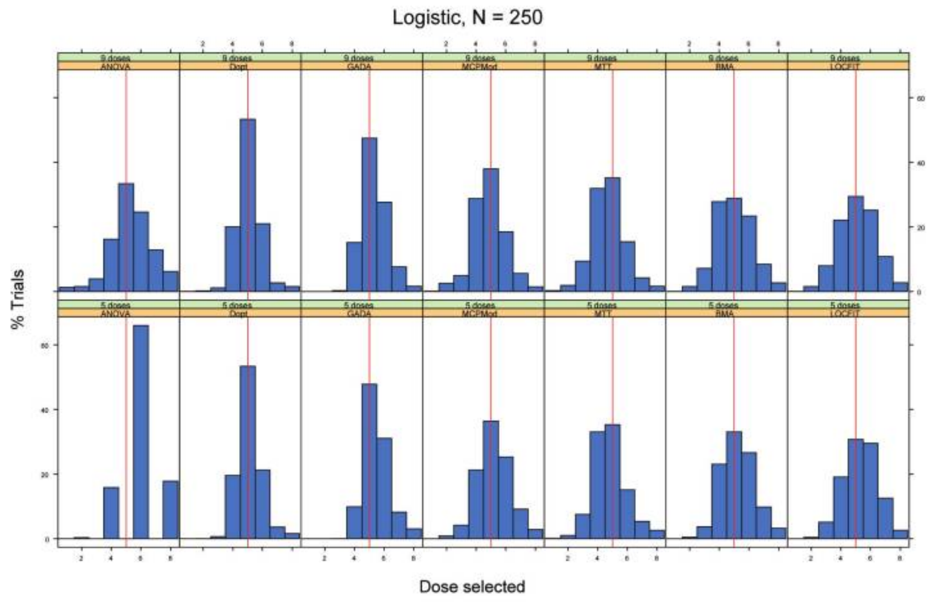

Figure 8. Distribution of selected dose with logistic dose-response. This figure illustrates the accuracy of the dose selection criteria for the logistic dose-response model. The red vertical line in each panel represents the dose with the highest overall probability of success.

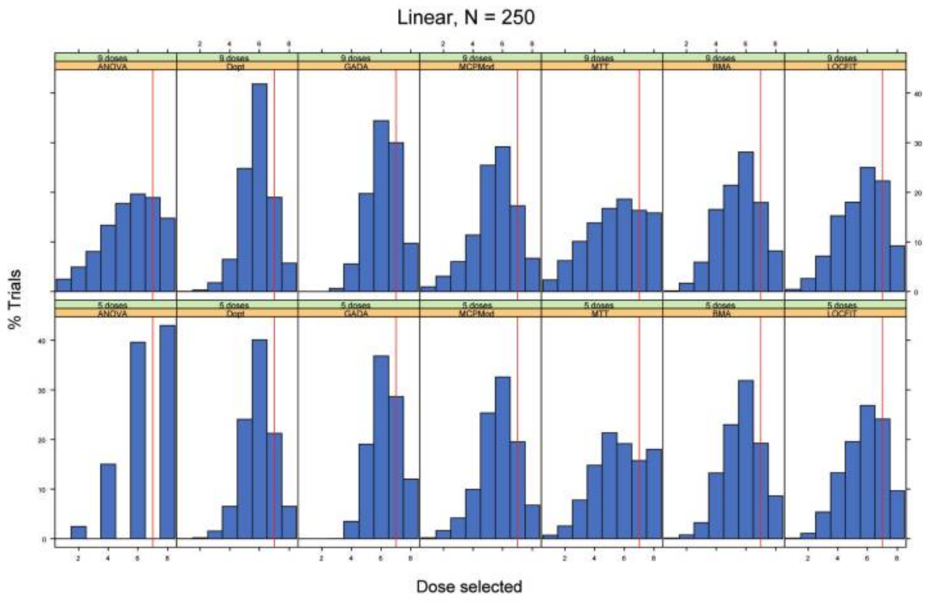

Figure 9. Distribution of selected dose with linear dose-response. This figure illustrates the accuracy of the dose selection criteria for the linear dose-response model. The red vertical line in each panel represents the dose with the highest overall probability of success.

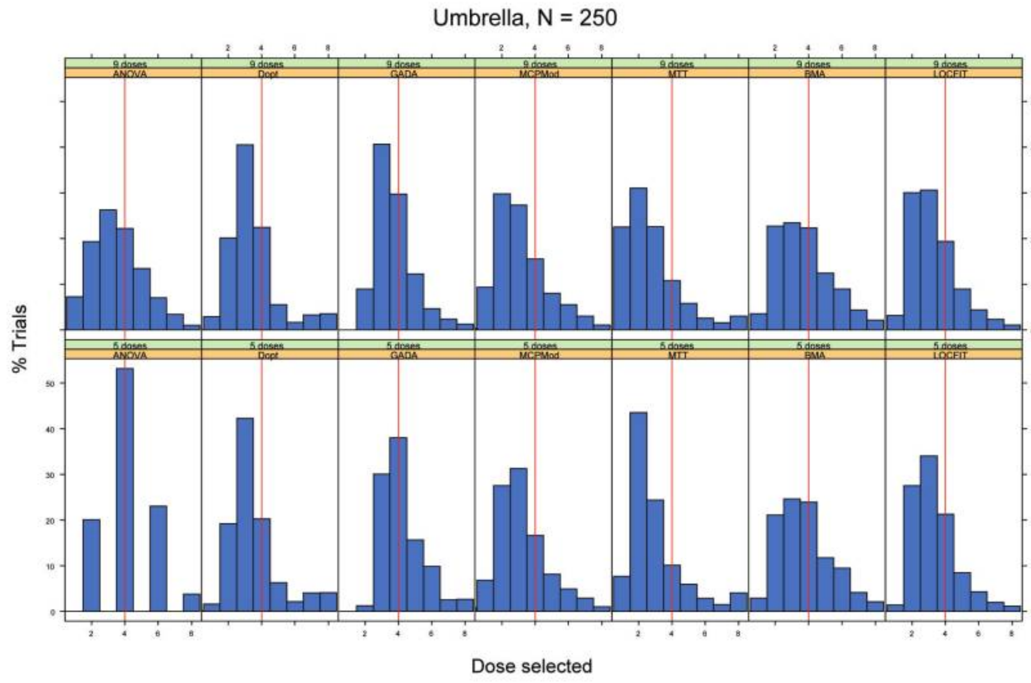

Figure 10. Distribution of selected dose with umbrella dose-response. This figure illustrates the accuracy of the dose selection criteria for the umbrella dose-response model. The red vertical line in each panel represents the dose with the highest overall probability of success.

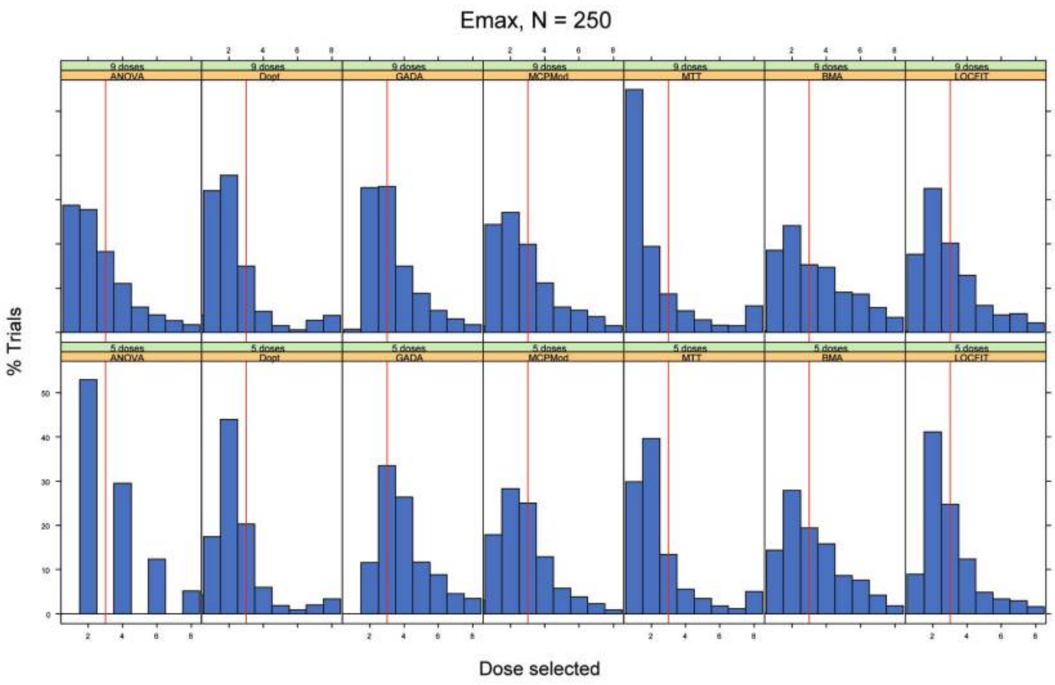

Figure 11. Distribution of selected dose with $E_{\max}$ dose-response. This figure illustrates the accuracy of the dose selection criteria for the $E_{\max}$ dose-response model. The red vertical line in each panel represents the dose with the highest overall probability of success.

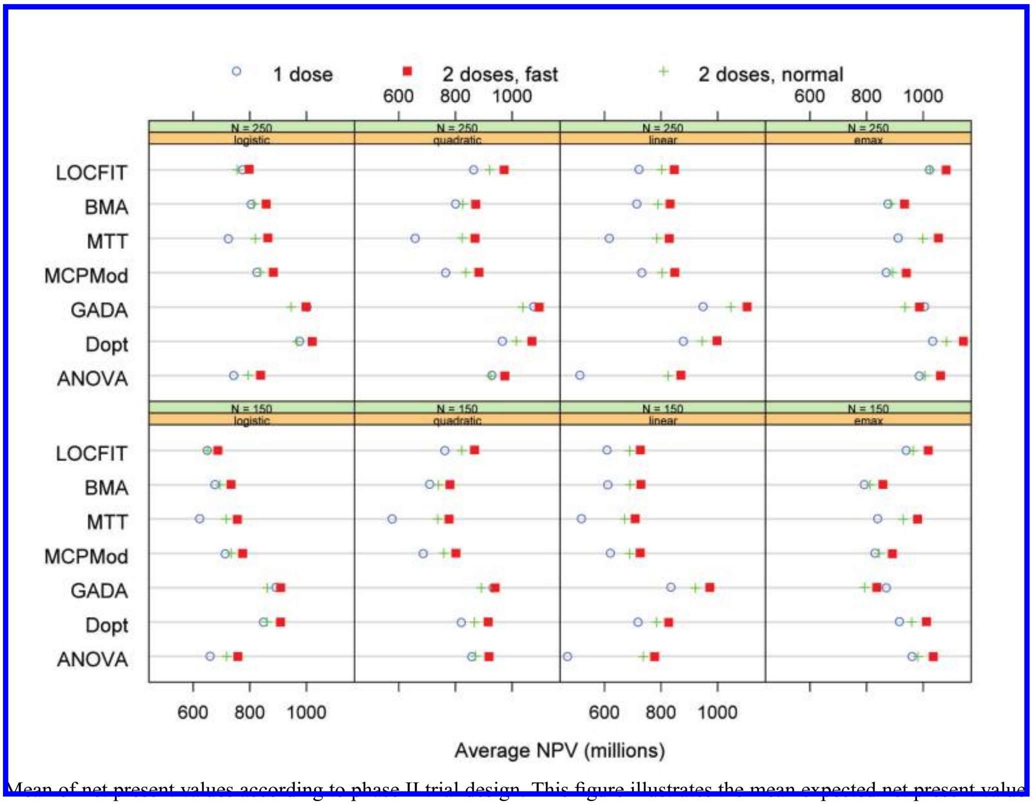

Figure 12. Mean of net present values according to phase II trial design. This figure illustrates the mean expected net present value according to the phase II design used, the underlying dose response, the number of doses carried forward into phase III, and the rapidity of enrollment in phase III.

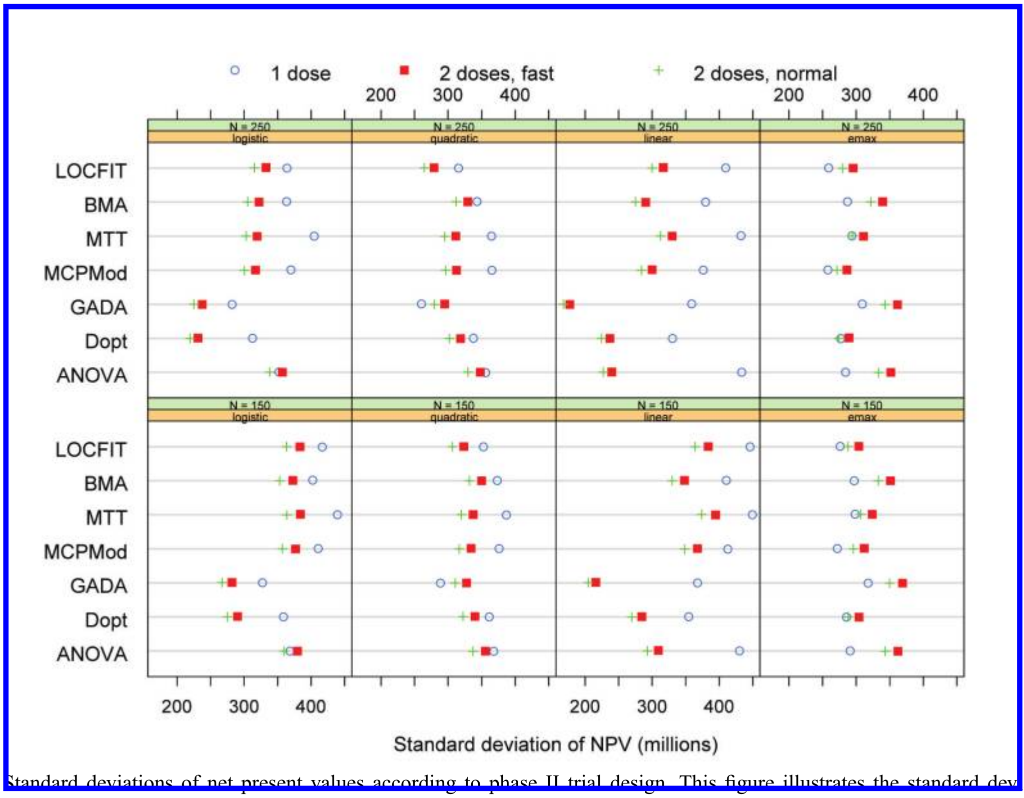

Figure 13. Standard deviations of net present values according to phase II trial design. This figure illustrates the standard deviations of the expected net present value according to the phase II design used, the underlying dose response, the number of doses carried forward into phase III, and the rapidity of enrollment in phase III.

was between 35% and 65% (35% to 60% for the phase III design with one dose). It is noteworthy that the reported industry failure rate is somewhere in the middle of that range. We know that our assumed dose-response profiles allowed for a maximum differentiation from placebo of 1.65 points, which is larger than the MCMD that we used for our sample size calculations. So why was the rate of failure so high? There are several likely reasons for this that are sometimes ignored, sometimes overlooked, and rarely addressed in study planning:

- A successful regulatory submission usually requires two pivotal trials; however, power calculations are traditionally performed at the individual study level and not at the program level.
- Individual studies and entire drug development programs may fail because of the observance of unacceptable toxicity.
- It is likely that suboptimal dose(s) are frequently selected for evaluation in phase III.

There is a known asymmetry in all stages of drug development, in that more attention is given to the evaluation of efficacy than toxicity (O'Neill 2008). This is certainly true in the planning of most phase II dose-finding studies and, in this case, this asymmetry was reflected in our predefined dose-selection criteria. At the end of phase II, the size of clinical databases is generally limited, and thus it is generally impossible to draw firm conclusions regarding product toxicity. Nonetheless, one should still make the best use of the available data in selecting the dose or doses to be carried forward into phase III. Some possible approaches to consider are:

- Selecting the maximum dose beyond which no further beneficial effect is seen.
- Choosing the highest dose which still appears to lack toxicity and shows a clinically relevant effect.
- Selecting the dose that yields the largest probability of success based on both its observed efficacy and toxicity profiles.
- Choosing a dose that maximizes a prespecified "utility" function.
- Selecting the minimum dose that delivers a predefined percentage of the maximum efficacy (e.g., the ED90).

4.2 Use of Simulation to Guide the Planning of Drug Development Programs

A key result of the current work is that the use of a traditional approach (e.g., ANOVA) to dose selection followed by bringing a single dose forward into phase III may be associated with a low probability of success and a markedly diminished expected net present value. While these results illustrate the potential for poor performance using traditional methods, they do not define a single best approach for all circumstances, although adaptive designs (and GADA in particular) appear to provide the most consistently well-performing approaches. Choosing the best approach for an individual program will generally require the use of simulations, similar to those performed in the current study, to determine the optimal structure and size for both the phase II dose-selection and phase III confirmatory trials. In conducting these simulations, a larger number of options for the number of doses to be included in phase II should be considered, as there may be a different optimal number of doses for any given method used. Based on our simulations, a larger number of doses should be included in a GADA phase II trial than with non-design-adaptive approaches.

4.3 Return on Investment in Phase II

While increasing the phase II sample size generally improves the probability of success in phase III only slightly, the calculation of expected net present value demonstrated that this investment is likely to pay off in the long run. A larger sample size in phase II increased both the probability of advancing into phase III, as well as the probability of success in phase III, while it only slightly increased the cost and slightly shortened the period of exclusivity.

The probability of success is consistently much higher in phase III designs with two active doses than those with only one active dose and this results in an increase in the expected net present value. After a traditional phase II dose-finding design, there is a big improvement in the probability of success if two doses are studied in phase III and the resulting change in expected net present value is positive. There is a smaller improvement after the conduct of a GADA dose-finding trial. This suggests that, by using an adaptive design in phase II, one may reduce but not eliminate the need for a larger investment in phase III. These conclusions, however, are based on simulations and calculations conducted under a specific set of assumptions, and we recommend that program-specific simulations are used to help with decision making for any given development program.

5. Conclusions and Recommendations

Based on the specific dose-response relationships modeled and dose-finding approaches studied, we found that adaptive phase II dose-finding trial designs, and GADA in particular, generally outperformed other designs across a variety of the dose-response, sample size, number of doses studies, or the number of doses advanced into phase III. Thus, we believe the development and evaluation of new phase II design-adaptive dose-finding designs should be a key area of future development. Simulation should play a key role in the development and evaluation of new approaches, as well as in the selection of strategies for specific development programs.

It is noteworthy that the dose-selection criteria we applied often failed to select the optimal dose for phase III. Thus, we believe greater emphasis should be placed on selecting a dose or multiple doses that have the best chance for approval based on both efficacy *and* a lack of toxicity and that dose-selection criteria should directly reflect this goal.

In the cases that we modeled, the probability of success and the period of exclusivity have more impact on the average expected net present value than the cost of the program. This is likely to be the case for any indication from which large revenues are expected. The investment into a particular program, however, usually comes from a limited source of money and it is important to analyze investments affecting multiple programs at the portfolio level.

Acknowledgments

The authors thank Frank Bretz, Vlad Dragalin, and Alex Dmitrienko for helpful advice during the planning and execution of this study.

[Received December 2008. Revised July 2009.]

REFERENCES

Berry, D. A., Müller, P., Grieve, A.P., Smith, M., Parke, T., Blazek, R., Mitchard, N., and Krams, M. (2001), "Adaptive Bayesian Designs for Dose-Ranging Drug Trials," in *Case Studies in Bayesian Statistics V* , Eds: C. Gatsonis, R. E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli, and M. West, New York: Springer-Verlag, pp. 99–181. 471

Bornkamp, B., Bretz, F., Dmitrienko, A., Enas, G., Gaydos, B., et al. (2007), "Innovative Approaches for Designing and Analyzing Adaptive Dose-Ranging Trials," *Journal of Biopharmaceutical Statistics*, 17, 965–995. 470, 471, 472, 473, 474, 475

Bretz, F., Pinheiro, J., and Branson, M. (2005), "Combining Multiple Comparisons and Modeling Techniques in Dose–Response Studies," *Biometrics*, 61, 738–748. 472

Dragalin, V., and Fedorov, V. (2006), "Adaptive Designs for Dose-Finding Based on Efficacy–Toxicity Response," *Journal of Statistical Planning and Inference*, 136, 1800–1823. 471

Dunnett, C.W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, 50, 1096–1121. 472, 474, 475

Gallagher, E.J., Liebman, M., and Bijur, P.E. (2001), "Prospective Validation of Clinically Important Changes in Pain Severity Measured on a Visual Analog Scale," *Annals of Emergency Medicine*, 38, 633–638. 472

Grieve, A.P., and Krams, M. (2005), "ASTIN: A Bayesian Adaptive Dose-Response Trial in Acute Stroke," *Clinical Trials*, 2, 340–351. 470, 471

Hemmings, R. (2007), Discussion of "The White Paper of the PhRMA Working Group on Adaptive Dose-Ranging Designs," *Journal of Biopharmaceutical Statistics*, 17, 1021–1027. 471

Ivanova, A., Bolognese, J.A., and Perevozskaya, I. (2008), "Adaptive Dose Finding Based on t-Statistic for Dose-Response Trials," *Statistics in Medicine*, 27, 1581–1592. 470

O'Neill, R.T. (2008), "A Perspective on Characterizing Benefits and Risks Derived from Clinical Trials: Can We do More?" *Drug Information Journal*, 42, 235–245. 485

About the Authors

Zoran Antonijevic is Senior Director, Center for Statistics in Drug Development, Innovation, Quintiles, Inc., Morrisville, NC 27560 (E-mail for correspondence: *Zoran.Antonijevic@Quintiles.com*). José Pinheiro is Senior Director, Adaptive Designs, Johnson & Johnson Pharmaceutical Research and Development, 920 Rt 202 South, P.O. Box 300, Raritan, NJ 08869 (E-mail: *jpinhei1@its.jnj.com*). Parvin Fardipour is Senior Director, Adaptive Designs, Pfizer, 500 Arcola Road, Collegeville, PA 19426 (E-mail: *Parvin.Fardipour@pfizer.com*). Roger J. Lewis is Vice Chair, Academic Affairs, Department of Emergency Medicine, Harbor-UCLA Medical Center, 1000 W. Carson Street, Box 21, Torrance, CA 90502 (E-mail: *roger@emedharbor.edu*).

This article has been cited by:

1. Yuki Ando, Akihiro Hirakawa. 2010. Discussion of "Adaptive and Model-Based Dose-Ranging Trials: Quantitative Evaluation and Recommendations"Discussion of "Adaptive and Model-Based Dose-Ranging Trials: Quantitative Evaluation and Recommendations". *Statistics in Biopharmaceutical Research* **2**:4, 462-465. [Citation] [PDF] [PDF Plus]
2. Sue-Jane Wang. 2010. The Bias Issue Under the Complete Null With Response Adaptive Randomization: Commentary on "Adaptive and Model-Based Dose-Ranging Trials: Quantitative Evaluation and Recommendation"The Bias Issue Under the Complete Null With Response Adaptive Randomization: Commentary on "Adaptive and Model-Based Dose-Ranging Trials: Quantitative Evaluation and Recommendation". *Statistics in Biopharmaceutical Research* **2**:4, 458-461. [Citation] [PDF] [PDF Plus]
3. Vladimir Dragalin, Björn Bornkamp, Frank Bretz, Frank Miller, S. Krishna Padmanabhan, Nitin Patel, Inna Perevozskaya, José Pinheiro, Jonathan R. Smith. 2010. A Simulation Study to Compare New Adaptive Dose–Ranging DesignsA Simulation Study to Compare New Adaptive Dose–Ranging Designs. *Statistics in Biopharmaceutical Research* **2**:4, 487-512. [Abstract] [PDF] [PDF Plus]
4. José Pinheiro, Frederic Sax, Zoran Antonijevic, Björn Bornkamp, Frank Bretz, Christy Chuang-Stein, Vladimir Dragalin, Parvin Fardipour, Paul Gallo, William Gillespie, Chyi-Hung Hsu, Frank Miller, S. Krishna Padmanabhan, Nitin Patel, Inna Perevozskaya, Amit Roy, Ashish Sanil, Jonathan R. Smith. 2010. Adaptive and Model-Based Dose-Ranging Trials: Quantitative Evaluation and Recommendations. White Paper of the PhRMA Working Group on Adaptive Dose-Ranging StudiesAdaptive and Model-Based Dose-Ranging Trials: Quantitative Evaluation and Recommendations. White Paper of the PhRMA Working Group on Adaptive Dose-Ranging Studies. *Statistics in Biopharmaceutical Research* **2**:4, 435-454. [Abstract] [PDF] [PDF Plus]