White Paper of the PhRMA Working Group on Adaptive Dose-Ranging Studies

Innovative Approaches for Designing and Analyzing Adaptive Dose-Ranging Trials

Björn Bornkamp, *Univ. of Dortmund*; Frank Bretz, *Novartis*; Alex Dmitrienko, *Eli Lilly*; Greg Enas, *Eli Lilly*; Brenda Gaydos, *Eli Lilly*; Chyi-Hung Hsu, *Pfizer*; Franz König, *Medical Univ. of Vienna*; Michael Krams, *Wyeth*; Qing Liu, *Johnson & Johnson*; Beat Neuenschwander, *Novartis*; Tom Parke, *Tessella*; José Pinheiro, *Novartis*; Amit Roy, *Bristol-Myers Squibb*; Rick Sax, *AstraZeneca*; Frank Shen, *Bristol-Myers Squibb*

1. BACKGROUND

In the spring of 2005, the Pharmaceutical Innovation Steering Committee (PISC) of PhRMA formed several working groups to look into different drivers of the decreasing success rates observed in drug development programs across the pharmaceutical industry, identified in a previous survey conducted by a consulting group. Among those was the Adaptive Dose-Ranging Designs (initially called Rolling Dose Studies) working group. The objectives of this group are to develop new and evaluate existing adaptive dose-ranging methods, and to produce policy recommendations for regulatory agencies on their use in clinical drug development. This paper summarizes the work of the group, including the results and conclusions of a comprehensive simulation study, and puts forward recommendations on how to improve dose ranging in clinical development, including, but not limited to, the use of adaptive dose-ranging methods.

2. INTRODUCTION

Selection of a dose (or doses) to carry into confirmatory Phase III trials is among the most difficult decisions that need to be made during drug development. Although the exact numbers are not known, it is thought that the high attrition rate which continues to plague the pharmaceutical industry in Phase III studies may, in part, be due to inadequate dose selection – doses that are too low to achieve adequate benefit, as well as doses that are too high and lead to dose-related adverse events in the population. There is also evidence that, even after registration, dose-adjustments in the label continue to be required with some frequency.

Until recently, the process of dose selection in a clinical program tended to follow a "traditional" pattern: When a new candidate drug was first studied in man, single ascending doses were used, followed by multiple ascending doses in healthy volunteers. The goal was to empirically establish both a "no effect" dose and a dose where subjects began to experience "symptom-limiting" adverse events (a "maximum tolerated dose"). This was followed by empiric testing of doses in small predefined patient cohorts, looking both for differences in the pharmacokinetic profile in patients (compared to healthy volunteers) and evidence that the drug had an effect on the disease being studied, often called a proof-of-concept (PoC) study. Inevitably, a point would come in the development of the drug where a critical decision had to be made as to gathering more

safety and efficacy data related to dosing in a larger patient cohort ("Phase IIb study"), or taking the plunge into a "development for launch (Phase III)" program to register the drug. At times, only a small amount of clinically-derived scientific data are available to justify the dose selection..

As we move into an era in which drug development is guided increasingly by the availability of biomarkers and modeling-and-simulation activities (from animal data to humans, from biomarkers to disease state and from disease state to clinical outcome), dosing decisions can be guided by the availability of prior information. This changes the paradigm of dose selection from one based on a purely empiric approach ("try a dose and let's see what happens") to an approach that is based more on verification of a derived model. The availability of "prior" information also allows one to prospectively specify adaptations to take place during a dose-ranging trial; such pre-specified adaptations allow modifications to the original study design as dose-response information accrues. This is known as adaptive dose-ranging and is the subject of this paper.

Adaptive dose-ranging can have a number of objectives. For example, it can be used to establish the overall dose-response relationship for an efficacy parameter, estimate the therapeutic window and help with the selection of a single target dose (*e.g.,* a dose that delivers a prescribed level of efficacy and is adequately safe). Other applications of adaptive dose-ranging designs might be related to specific indications; for example, an oncology study might use an adaptive approach to find a maximum tolerated dose.

Adaptive dose-ranging approaches can encompass a variety of design modifications, including adding doses, dropping doses, changing sample sizes and unbalanced randomization (randomizing more patients to doses of interest), pre-specified in the study protocol to occur as information accrues during the course of the trial. Adaptations to the analysis of dose-response data can also be performed even if fixed randomization schedules are used, again based on information that becomes available during the course of a study. Use of such adaptations could substantially enhance the efficiency and effectiveness of dose-selection in clinical programs – increasing the amount of information available to allow for good selection decisions, while decreasing the time and cost of running these studies.

In addition to adaptive dose-ranging being used to address a number of different objectives and encompassing a number of different approaches, there are also a number of different statistical methodologies that have recently been developed to allow adaptations to take place without compromising the statistical integrity of the trial (*e.g.,* preserving the Type I error rate). This paper explores some of these statistical strategies and attempts to assess the performance of these approaches. A variety of methodologies are considered, from more traditional frequentist approaches to flexible Bayesian approaches. Using a common format and set of parameters across all approaches (pre-specified scenarios for dose selection; the context and parameters are described in Section 4), the strengths and weaknesses of each method are explored via a comprehensive simulation study, not only with respect to the robustness of information obtained, but also the ability of the approach to learn from the data and adapt to emerging trends. In doing so, this paper provides insight into understanding which adaptive dose-ranging strategies might be most useful for establishing the presence of dose-response, estimating the dose-response relationship and selecting a target dose.

Since the focus of this paper is on statistical methodologies for adaptive dose-ranging, general considerations on the use of adaptive designs are not discussed here. A thorough discussion of these can be found in the special issues of the *Biometrical Journal* (Vol. 48, No. 4, 2006) and the *Drug Information Journal* (Vol. 40, No. 4, 2006). Still, certain assumptions are worth mentioning. First, the scenarios and methods considered in this paper are purely for dose-ranging purposes; thus, regulatory concerns often associated with adaptive trials in the confirmatory setting (Phase III) are less of an issue. Nevertheless, to ensure the validity of the trial results, there has to be an implicit assumption that demographic and other relevant characteristics of the patients enrolled in the study remain relatively constant over time. Second, it is assumed that the efficacy response being measured is available in a sufficiently rapid time frame (relative to the enrollment period and the duration of the study) to allow for meaningful adaptations to occur (the scenarios described in this paper involve a visual analogue scale measurement of neuropathic pain at a 6 week period). Lastly, it is assumed that the transmission of relevant information to the data analysis group is sufficiently rapid (*e.g.*, instant data capture and transmittal) to allow adaptations to occur according to the prescribed methodology. These considerations are pre-requisites for any adaptive dose-ranging study and thus must be taken into consideration before embarking on any such program.

The rest of the paper is organized as follows. The dose-ranging methods utilized in the simulation study are described in Section 3. Section 4 is dedicated to the simulation study itself, describing the different scenarios and assumptions used and presenting key summary results. Logistical and regulatory considerations are discussed in Section 5, with the working group's recommendations being presented in Section 6.

3. STATISTICAL METHODOLOGY

This section provides an overview of traditional and adaptive approaches to designing and analyzing dose-ranging studies. Two classes of adaptive procedures will be described below. The first class includes methods that enable clinical trials researchers to modify certain elements of the study design based on the data collected in the trial. For example, based on the review of interim safety and tolerability data, the trial's sponsor can decide to drop one or more doses or reduce the number of patients assigned to these doses. This adaptation can be performed in a continuous manner, *i.e.*, the design can be updated prior to the enrolment of each new patient. Alternatively, one can consider group-sequential adaptive strategies and perform design modifications based on responses from cohorts of patients. This class of methods will be termed the *design-focused adaptive approaches*.

The other class of approaches (*analysis-focused adaptive approaches*) relies on a fixed study design, that is, no design modifications are performed during the study. In this case, the trial's sponsor focuses on selecting the most appropriate data analysis method. The choice of the "best" method is adaptive in the sense that it is driven by the data collected in the trial. To illustrate this approach, consider the case when the sponsor examines the data to find a model for the dose-response relationship that improves their ability to detect the dose-related drug effect and characterize the dose-response function.

In each of the two classes of approaches, adaptation is aimed at optimizing the assessment of the dose-response relationship and narrowing down the set of doses from which to select the estimated target dose, for example, by identifying the doses that

provide a pre-specified level of improvement. The statistical approaches introduced in the section are listed below.

- **Traditional approach (ANOVA):** this approach relies on the classical analysis of variance (ANOVA) methods. The data analysis procedures are pre-specified and no adaptive elements are employed within the traditional framework.
- **Design-focused adaptive approaches**
  - **General adaptive dose allocation approach (GADA)** utilizes Bayesian modeling to identify an appropriate dose for each new patient based on partial and complete responses from the previously enrolled patients (design modifications are carried out in a continuous manner).
  - **D-optimal response-adaptive approach (Dopt)** relies on the D-optimality criterion to select the patient allocation scheme that provides most information about the dose-response relationship given the interim data. The method is applied in a group-sequential manner.
- **Analysis-focused adaptive approaches**
  - **Multiple comparison procedures–modeling approach (MCP-Mod)** combines elements of multiple testing and modeling to select an appropriate dose-response model and subsequently estimate the target dose.
  - **Bayesian model-averaging approach (BMA)** utilizes Bayesian methods to produce weights used to combine information obtained from a set of pre-specified dose-response models
  - **Multiple trend test approach (MTT)** utilizes three sigmoid Emax models to describe a range of possible dose-response curves and improve the clinical researchers' ability to identify a target dose.
  - **Nonparametric dose-response modeling approach (LOCFIT)** relies on nonparametric regression methods, based on local polynomials (LOESS), to allow greater flexibility in the estimation of the dose response.

3.1 Traditional approach (ANOVA)

The traditional ANOVA approach uses contrast statistics comparing the dose levels (*e.g.*, differences between active doses and placebo) to determine the existence of a dose-response (PoC) and, if so, to select a target dose. Multiplicity adjustments (e.g., Dunnett, 1955) are typically used to preserve the family-wise Type I error rate associated with the contrast tests. These methods can optionally be combined with post-hoc modeling of the dose-response and with strategies for incorporating clinical relevance in dose selection.

The particular ANOVA approach used in the simulation study described in Section 4 consists of an initial one-sided Dunnett multiple comparison procedure to test each of the active doses against placebo. If at least one of the doses is statistically significant (under Dunnett's multiplicity adjustment), dose-response is established. The target dose is then estimated as the smallest statistically significant dose which has an average effect that is clinically relevant (according to a pre-specified value of clinical relevance), provided at least one dose meet both criteria. If a target dose can be estimated, the final step of the approach consists in estimating a dose-response model. Three candidate dose-response models (linear, quadratic, and logistic) are fitted to the data and the Akaike Information Criterion (AIC) is used to select the best model, which is then used for predictions, etc.

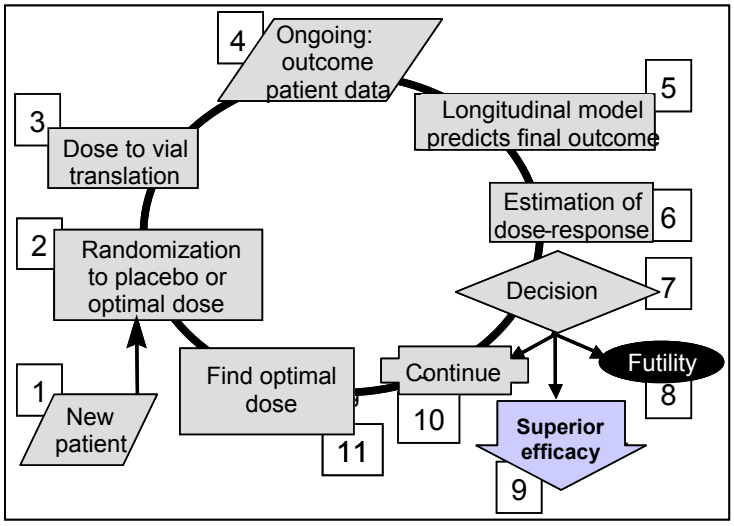
3.2 Bayesian adaptive dose allocation approach (GADA)

This design-focused adaptive approach relies on Bayesian dose-response modeling and employs a decision theoretic framework to determine the most informative dose to administer to each new subject (Berry et al, 2002). This method generalizes the approach implemented in the ASTIN trial (Krams et al, 2003), a Phase II study in stroke patients.

The key elements of this adaptive approach are depicted in Figure 3.1. A new subject enters the study [1] and is randomized to either placebo or an active dose based on the optimal allocation rule [2]. The dose allocation decision is then converted to an appropriate dosing instruction [3] that, for example, gives a combination of tablets that results in the required dosing level. As the trial proceeds, interim and final results from subjects are gathered and sent to the central system (as soon as they are available) in addition to CRF data [4]. Subjects with interim results but as yet no final results can have final results imputed by a longitudinal model [5] (this facility was disabled in the simulation study described in Section 4). The final (and imputed) results are then used to update the estimated dose response model [6]. Based on the posterior probability of the response at the target dose (*e.g.*, ED95 or the dose with a clinically relevant response), a decision [7] is made whether to stop the trial for futility [8], superior efficacy [9] or continue with the dose finding stage [10]. The optimal dose for the next patient for the next subject is chosen by simulating the effects of randomizing the next subject to each of the possible doses and finding the one that minimizes the variance of a parameter of interest (typically this is the variance in the response at the target dose).

The outlined dose allocation method uses the normal dynamic linear model (NDLM) (West and Harrison, 1997) to model the dose-response curve, combined with a conventional Gamma distribution model for the inter-subject variance. The NDLM provides a non-parametric, non-monotonic model for smoothing the dose-response function that can be updated analytically. A non-informative prior is used for the dose-response relationship. Key characteristics of the adaptive design (*e.g.*, Type I and Type II error rates) are estimated by simulation.

Figure 3.1. Bayesian adaptive dose allocation approach.



3.3 D-optimal response-adaptive approach (Dopt)

In the context of design-focused adaptive approaches, clinical trial researchers are frequently interested in updating patient allocation to the active doses and placebo in order to improve the estimation of various aspects of the dose-response relationship. This includes estimation of a single target dose (*e.g.*, a dose that delivers a pre-specified level of efficacy such as ED90), estimation of the therapeutic window (efficacious doses with an acceptable safety profile) or the entire dose-response curve.

The D-optimal approach focuses on minimizing the variance of model parameters describing the dose-response function, including the baseline effect, slope and maximum effect. In order to implement this method, one needs to choose a rich family of models (*e.g.*, a logistic model with a polynomial trend) that can provide a good fit to dose-response curves likely to be encountered in a particular setting (both monotone and non-monotone). Given this family of models, the D-optimal patient allocation algorithm is easy to apply in a group sequential manner (patients are enrolled in groups and an interim analysis is performed after each group has completed the study) as described below.

The first cohort of patients is randomized to the active doses and placebo using any pre-specified patient allocation scheme, *e.g.*, equal allocation. At the first interim look, a sigmoid or other appropriate (non-linear) model is fitted to the data to estimate the dose-response curve. Based on the obtained model, a patient allocation scheme that maximizes the amount of information about the overall dose-response relationship is found. In most cases, it is reasonable to keep the proportion of patients assigned to placebo constant throughout the study to facilitate the dose-placebo comparisons.

Patients in the second cohort are randomly allocated to the active doses and placebo according to the derived optimal allocation scheme and, at the second interim look, a model is fitted to the data to update the estimate of the dose-response function. The computed dose-response function is again utilized to define an optimal allocation scheme for the next cohort. This process is repeated until the total sample size is reached or a pre-specified futility rule is met.

At the end of the study, the estimation of the dose-response relationship as well as optimal dose selection can be performed using pre-determined or data-driven methods. Due to the response-adaptive nature of the described approach, the final analysis of the dose-response relationship will be more informative than the analysis performed in the traditional fixed-design setting.

3.4 Multiple comparison-modeling approach (MCP-Mod)

Multiple comparison procedures (MCP) and modeling techniques have traditionally been the two main approaches used to design and analyze dose-ranging studies in drug development. Bretz, Pinheiro and Branson (2005) proposed a hybrid methodology (termed the *multiple comparison-modeling or MCP-Mod approach*) which combines aspects of MCP and modeling into a unified strategy for dose-ranging studies.

This methodology allows both the statistical testing of evidence of dose-response (PoC), as well as the estimation of target doses to be used in confirmatory studies. Instead of pre-specifying a single dose-response model, the MCP-Mod approach uses a set of

candidate models covering a suitable range of dose-response shapes. Each of the models in the candidate set is assessed using appropriately defined contrast tests and employing MCP techniques to preserve the family-wise error rate (FWER). PoC is established when at least one of the model contrast tests is significant. Otherwise, the procedure stops and concludes that there is no sufficient evidence of a dose-response relationship in the study.

After PoC has been established, the best model is selected from the statistically significant models in the candidate set. The selection of the best model can be based on the minimum *p*-value of the test statistics or some other relevant model selection criteria such as the AIC or the Bayesian Information Criterion (BIC). The selected dose-response model is then employed to estimate target doses using inverse regression techniques and possibly incorporating information on clinically relevant effects. The precision of the estimated doses can be assessed using, for example, bootstrap methods.

The MCP-Mod method combines the advantages of multiple comparisons and modeling. The first step provides robustness to model misspecification and places the associated statistical uncertainty in a hypothesis testing context, while the second step (dose-response modeling) provides greater flexibility and efficiency in estimating target doses.

3.5 Bayesian model-averaging approach (BMA)

The Bayesian approach to dose ranging in clinical trials has a modeling and a decision-making component. The former necessitates a flexible class of dose-response models (parametric or nonparametric), whereas the latter relies on decisions that are based on appropriate inferential summaries and can be either informal (posterior- and/or predictive-based) or formal, *i.e.*, fully decision-analytic using utilities. Dose-response modeling can be done in various ways, ranging from simple models with a small number of parameters (typical for small Phase I trials) to high-dimensional or non-parametric models. Bayesian model averaging is an intermediate strategy that tries to avoid the dangers of under- or over-fitting: the basis is a set of relatively simple dose-response models with corresponding parameters. Then, starting from prior model probabilities ("weights"), as well as prior distributions on the model-specific parameters, standard Bayesian inference leads to posterior updates of the unknown quantities (model weights and model parameters). The former are usually more difficult to obtain even with the help of the Markov chain-Monte Carlo (MCMC) approach. Therefore, in the case of non-informative priors, the updates are often approximated by the BIC. Bayesian model averaging is conceptually straightforward and can be seen as a special case of a hierarchical model. The approach is well-suited for situations where the quantity of interest is model-independent, such as in dose-ranging studies where the objective is to find a dose fulfilling a certain pre-specified criterion. Bayesian model averaging generalizes model selection strategies and has the advantage of weighting the candidate models in an appropriate (data-dependent) way.

A simple informal Bayesian model averaging approach based on a set of normal linear models allowing for analytic posterior updates is used in this paper. This choice was mainly dictated by the fact that MCMC-based posterior inferences for non-conjugate models would have made simulations computationally infeasible. The approach is informal in that it only uses posterior summaries as a basis for dose selections. Depending on the situation at hand and available external input, the approach can be

tailored accordingly by selecting a more appropriate model basis, including prior information, and extending dose selection to a fully decision-oriented framework.

3.6 Multiple trend test approach (MTT)

To reflect most dose-response curves one would encounter in dose-ranging studies, it is sensible to select, from a class of sigmoid Emax models, two dose-response (upper and lower) curves that capture a range of likely dose-response curves. A dose-response curve in the middle of the range is also selected from the class of Emax models so that, for a given sample size, the power of the resulting triple-trend test is minimized. The sample size is chosen to achieve pre-specified minimum power, for example, 95%. This yields a procedure with robust power for a broad spectrum of underlying dose-response curves.

For the given class of Emax models, the parameters, and the associated dose-response curve, are estimated by maximum likelihood. For an estimated curve, a reverse calculation yields an estimate of the target dose, which is the dose corresponding to a target effect size. However, a target dose is not an underlying characteristic of the model, and therefore, it may not be identifiable for a given estimated dose-response curve.

The outlined approach extends the work of Tukey, Ciminera and Heyse (1985) and Capizzi et al (1992). Tukey, Ciminera and Heyse (1985) introduced three-trend tests for carcinogenicity and toxicology animal studies without a multiplicity adjustment and Capizzi et al (1992) proposed to use three-trend tests for efficacy analysis, where correlations between three trend test statistics are explicitly incorporated to control the Type I error rate.

3.7 Nonparametric dose-response modeling approach (LOCFIT)

This method relies on model-free testing techniques to assess a possible dose-response effect. Nonparametric regression techniques are used for target dose estimation as they can model virtually any smooth dose-response shape without the need to pre-specify a parametric dose-response model.

The dose-response effect is assessed using a multiple contrast test. To cover a broad range of potential dose response shapes, the method relies on five contrast tests capturing the concave, convex, sigmoid, linear and umbrella model shapes (see Stewart and Ruberg (2000) for more information on multiple contrast tests).

For the dose estimation step we utilized local quadratic regression techniques (Loader, 1999) using a Gaussian kernel and a global bandwidth. The bandwidth was selected by minimizing the generalized cross-validation score. The *locfit* package in R was used for the implementation of the procedure.

4. SIMULATION STUDY TO EVALUATE PERFORMANCE OF METHODS

One of the key goals of the Adaptive Dose-Ranging Studies working group is to provide an assessment of the relative performance of the traditional and novel dose-ranging methods described in Section 3, leading to recommendations on their use in practice. To allow a direct quantitative comparison of the described methods, under the same conditions and using the same performance metrics, a comprehensive simulation study,

motivated by a real dose-ranging application and covering a wide range of practical scenarios, was undertaken. This section describes the design of this simulation study, including its assumptions and scenarios; the performance metrics used to evaluate different statistical operational characteristics of each method; and a graphical summary of the statistical performance of the methods, based on the simulation results.

To give practical motivation, a neuropathic pain dose-ranging study was used to provide context for the simulation study. *However, the assumptions and scenarios considered in the simulations, as well as the performance metrics used to summarize the results, are quite general, allowing the extension of results and conclusions to a wide range of trials.* In fact, we could have used any indication having dose-ranging trials with a normally distributed response and independent treatment groups, such as Type 2 diabetes (HbA1c endpoint), high blood pressure, dyslipidemia (LDL cholesterol endpoint), oncology (tumor size endpoint), etc. We expect the main conclusions of the simulation study to remain valid, at least qualitatively, to other types of response variables (e.g., binary, time-to-event) and designs (e.g., cross-over studies).

4.1 Design of simulation study

In total, 36 different scenarios were used in the simulations, corresponding to different combinations of dose designs (3), dose-response profiles (6), and total sample size (2).

Primary endpoint, distributional assumptions, and clinically relevant effect: following common practice in neuropathic pain studies, the primary endpoint is the *change from baseline* at 6 weeks in a *visual analog scale* (VAS) of pain. The VAS takes values between 0 (no pain) and 10 (highest pain) on a continuous scale and, for the purpose of the simulation, it is assumed to be normally distributed. Weekly VAS measurements are collected in each patient until the end of follow up, at 6 weeks. The between-patient standard deviation is set to 2.5 units and the within-patient standard deviation at 1.5 units. For simplicity, it is assumed that no patients drop out of the study. Letting VAS_k denote the VAS measurement at the k^{th} Week, k = 0, 1, …, 6 and with k = 0 representing baseline, the primary endpoint is defined as $y = VAS_6 - VAS_0$. From the assumptions above, y is normally distributed with variance $\sigma^2 = 2 \times 1.5^2 = 4.5$. Negative values of y give indication of efficacy in reducing the neuropathic pain. The clinically relevant effect is set to $e_{targ} = -1.3$ units (*i.e.*, an average reduction of at least 1.3 units from baseline).

Dose design scenarios: it is assumed that up to nine equally spaced doses can be utilized in the trial: 0, 1, …, 8 (note that these are not indices, but the actual dose values). Three different dose designs are considered in the simulations, to investigate the impact of number and spacing of doses on the performance of the methods:

- Five equally spaced doses: 0, 2, 4, 6, and 8.
- Seven unequally spaced doses: 0, 2, 3, 4, 5, 6, 8.
- All nine equally spaced doses: 0, 1, …, 8.

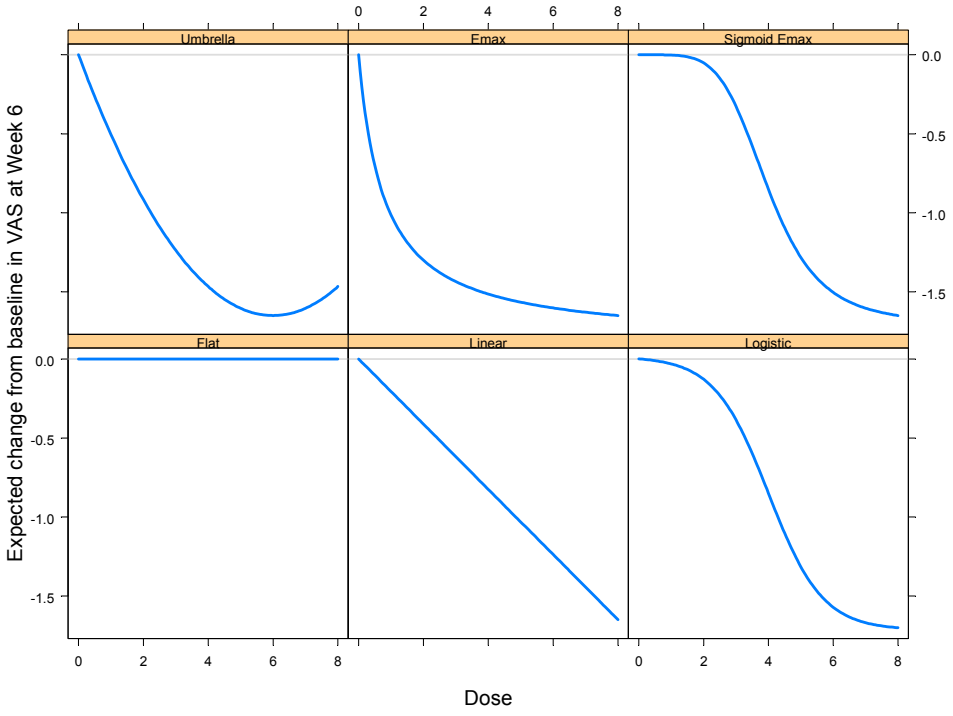
The scenarios above list the set of available study doses; the subset of doses actually utilized by a given method will depend on its specific characteristics and procedures.

Dose-response profiles: a total of six different dose-response profiles were used to simulate the primary endpoint, allowing the evaluation of the methods under a wide range of scenarios likely to be observed in clinical practice. A flat model was also included to evaluate the Type I error rate. In all models, the placebo effect was set to 0 points and, with the exception of the logistic model, the maximum effect within the observed dose range was set to -1.65 units.

- Flat: $y = \varepsilon$.
- Linear: $y = -(1.65/8)\, d + \varepsilon$.
- Logistic: $y = 0.015 - 1.73/\{1 + \exp[1.2(4 - d)]\} + \varepsilon$.
- Umbrella: $y = -(1.65/3)d + (1.65/36)d^2 + \varepsilon$.
- Emax: $y = -1.81\, d/(0.79 + d) + \varepsilon$.
- Sigmoid Emax: $y = -1.70\, d^5/(4^5 + d^5) + \varepsilon$.

The error terms ε are assumed to be independently normally distributed with mean zero and variance 4.5. Figure 4.1 presents the corresponding dose-response profiles.

Figure 4.1. Dose-response profiles used in simulation.



The dose response profiles in Figure 4.1 cover the most common DR shapes observed in dose-ranging studies. It is clear from Figure 1 that the dose response profiles corresponding to the logistic and sigmoid Emax models are very similar. As a consequence, the simulation results for the two models are quite similar and we will include only the logistic model summaries.

Sample size and number of simulated trials: two maximum total sample sizes are used in the simulations: 150 and 250 patients. These values are consistent with sample sizes commonly used in neuropathic pain Phase II trials. The total sample size N corresponds to the sum of the number of patients assigned to each dose. For example, under equal treatment allocation and with $N = 250$, a 5-dose design assigns 50 patients to each dose and a 9-dose design, about 28. Note that design-focused adaptive methods which have flexible treatment allocation rules, such as GADA and Dopt described in Section 2, will likely use different sample sizes per dose. The sample size N refers always to the maximum number of patients that can be used in the trial.

To adequately estimate the statistical operational characteristics of the various methods, a minimum of 5,000 simulated trials are used for each of the scenarios considered. For some of the less computationally intensive methods (*e.g.*, MCP-Mod and ANOVA), 10,000 simulated trials per scenario are used.

Number of adaptations and accrual period: these are only relevant for design-focused adaptive methods, and the latter only for methods using calendar data information (*e.g.*, GADA). The number of adaptations is method-specific and so no restrictions are placed on it. For simplicity, only one accrual period is used in the simulations: 16 weeks, corresponding to relatively fast accrual for a neuropathic pain indication.

4.2 Measuring performance of methods

The main goal of a Phase II program should be to ensure efficient learning of the dose-response profile to allow accurate and timely decision making with regard to moving into the confirmatory stage. To make the problem more concrete and easier to quantify, for the purpose of the simulation study the following specific goals were identified:

- **Detecting DR:** evaluate if there is evidence of activity associated with the drug, represented by a change in clinical response resulting from a change in dose (PoC);
- **Identifying clinical relevance:** if PoC is established, determine if a pre-defined *clinically relevant* response can be obtained within the observed dose range;
- **Selecting a target dose:** when the previous goal is met, select the dose to be brought into the confirmatory phase, the so-called *target dose*;
- **Estimating the dose response:** finally, estimate the dose-response profile within the observed dose range.

Performance metrics to quantify each of these goals are described below.

Detecting DR: each of the methods described in Section 3 includes a decision rule to determine whether the data provides sufficient evidence of dose response (DR) activity. This can be a simple hypothesis test (*e.g.,* ANOVA), or a Bayesian rule utilizing the posterior distribution of the model parameters (*e.g.*, GADA). The probability of identifying the presence of dose response, *Pr(DR)*, estimated as the percentage of simulated trials in which the decision rule concluded for DR activity, is used as the summary metric for this objective. Under a flat dose response (DR) model, *Pr(DR)* gives the Type I error rate and, under active dose-response profiles, it gives the power to make

the correct identification of DR. To allow adequate comparisons, a nominal Type I error rate of 5% was specified for all methods.

Identifying clinical relevance: *Pr(DR)* is based on the statistical evidence of a non-flat dose-response profile. It is, of course, possible to conclude that DR is present, but, nevertheless, that none of the observed doses is capable of producing at least the clinically relevant effect. All methods described in Section 2 implement decision rules for identifying clinical relevance within the dose range of the trial. The corresponding probability, *Pr(dose),* estimated as the percentage of simulated trials in which PoC was established and a dose with a clinically relevant effect was chosen, is used to summarize the performance of the methods with regard to this objective. By definition, *Pr(dose)* ≤ *Pr(DR)*. No nominal levels under the flat DR model are specified for *Pr(dose)*.

Selecting a target dose: in practice, the selection of the dose to bring into the confirmatory phase is based on a plurality of factors, including, but not restricted to, efficacy and safety outcomes in the Phase II trial(s). For the purpose of the simulation study, we simplify the problem and consider only a target efficacy result (the clinically relevant effect) to determine the dose to be selected. In this context, the target dose $d_{t\arg}$ is defined as the smallest dose which produces an effect greater than, or equal to, the clinically relevant target effect $e_{t\arg}$. Even though not explicit in the notation, it should be clear that $d_{t\arg}$ varies with the underlying DR model. So does the impact of deviations from $d_{t\arg}$ on the expected clinical effect. For example, in a DR model with a steep slope around $d_{t\arg}$, small deviations from the target dose could result in large deviations from the target effect. On the other hand, models with a shallow slope around $d_{t\arg}$ would be considerably less sensitive to deviations around the target dose. To account for this, we consider target effect intervals $I^{e}_{t\arg}(\eta) = e_{t\arg} \pm \eta\, e_{t\arg}$ (that is, being within ±100η% of the target effect) and their corresponding target dose intervals $I^{d}_{t\arg}(\eta)$.

Some of the methods described in Section 3 (*e.g.*, GADA and MCP-Mod) can provide model-based estimates for $d_{t\arg}$ on a continuous scale. Because of manufacturing restrictions which limit the availability of doses that can be used in practice, for the purpose of the simulation study the dose selection is restricted to the set $D = \{1, 2, \ldots, 8\}$. Therefore, estimated target doses resulting from any of the model-based methods are rounded to the nearest integer within the set *D*. Note that this may result in a dose not used in the trial being selected in the end as the target dose (*e.g.*, *d=3* being chosen in the 5-dose design). Table 3.1 lists the values of $d_{t\arg}$ and $I^{d}_{t\arg}(0.1)$, with their corresponding rounded values, for the five active DR models used in the simulations. Rounded values for $I^{d}_{t\arg}(0.1)$ were obtained taking into account clinical considerations and not necessarily correspond to the nearest element in *D*.

Table 4.1 Target doses and target dose intervals for DR models used in simulation.

Model	$d_{t\arg}$		$I^{d}_{t\arg}(0.1)$	
	actual	rounded	actual	rounded
Linear	6.30	6	(5.67, 6.93)	{6,7}
Logistic	4.96	5	(4.65, 5.35)	{5}
Umbrella	3.24	3	(2.76, 3.81)	{3,4}
Emax	2.00	2	(1.44, 2.95)	{2,3}
Sig-Emax	5.06	5	(4.68, 5.58)	{5}

Let $\hat{d}_{t\arg}$ represent the estimate of the target dose corresponding to a given method used in the simulations, appropriately rounded to the nearest element of *D*. The probability distribution of $\hat{d}_{t\arg}$, estimated from the simulated trials, provides a complete description of the performance of the estimate. Additional statistics are used to summarize the dose estimation performance of the various methods, with expectations and probabilities referring to the corresponding Monte Carlo distributions obtained in the simulations.

- Percentage bias: $pBias = 100[E(\hat{d}_{t\arg}) - d_{t\arg}] / d_{t\arg}$.
- Percentage absolute error: $pError = 100E \mid \hat{d}_{t\arg} - d_{t\arg} \mid / d_{t\arg}$.
- Probabilities of
  - under interval estimation: $P^{-}(\eta) = \Pr[\hat{d}_{t\arg} < d_{\min}(\eta)]$,
  - over interval estimation: $P^{+}(\eta) = \Pr[\hat{d}_{t\arg} > d_{\max}(\eta)]$, and
  - correct interval estimation: $P^{o}(\eta) = \Pr[d_{\min}(\eta) \le \hat{d}_{t\arg} \le d_{\max}(\eta)]$,

where $d_{\min}(\eta)$ and $d_{\max}(\eta)$ denote the endpoints of the target interval $I^{d}_{t\arg}(\eta)$, appropriately rounded to an element of *D*.

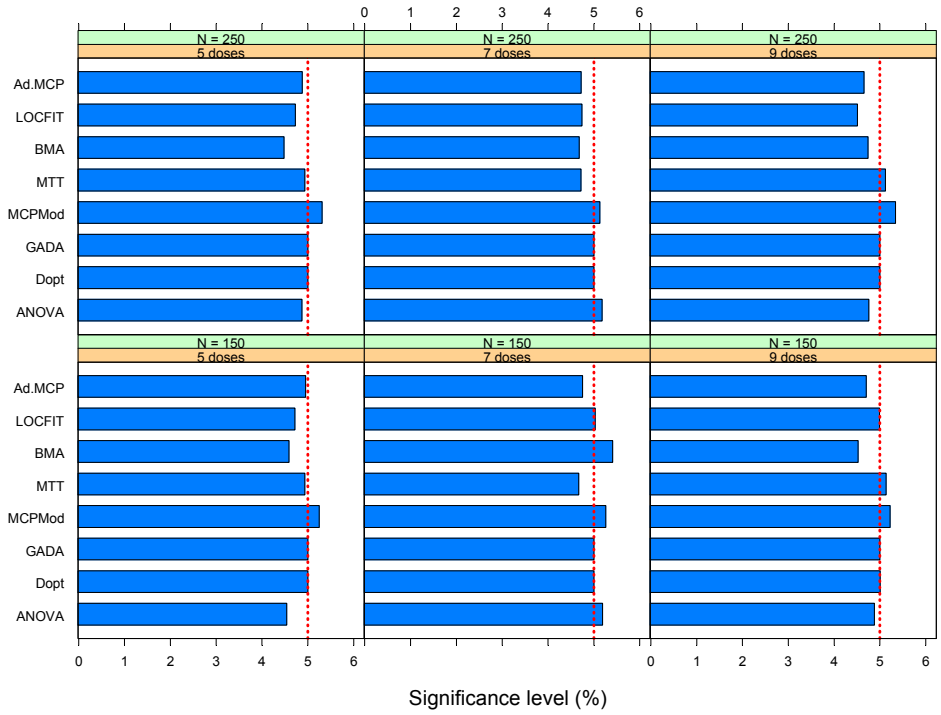
Estimating the dose response: proper estimation of the DR profile is relevant not only for the purpose of estimating target doses, but also for the appropriate labeling after approval. The average *absolute prediction error* (APE), calculated at the available doses (including placebo), is used as an overall measure of performance for DR estimation. Letting $\mu(d)$ denote the expected dose response at dose *d* and $\hat{\mu}(d)$ its prediction based on the estimated dose-response model, we define $APE = 1/9\sum_{d=0}^{8} E \mid \hat{\mu}(d) - \mu(d) \mid$. To make the summary statistic non-dimensional and interpretable as a percentage, we consider the *percent* APE (pAPE), defined as the percent value of APE relative to the (absolute) target effect, that is $pAPE = 100\,APE / \mid e_{t\arg} \mid$ (the absolute target effect does not provide an upper bound on the APE, so pAPE can take values greater than 100%). Quantiles of the prediction errors at each dose level (*i.e.,* $\hat{\mu}(d) - \mu(d)$) are also used in Section 3.3 to summarize the DR prediction performance of the various methods.

4.3 Simulation results

The results of the simulation study are summarized here using the performance metrics of Section 4.2. Because of the large number of scenarios and performance metrics, only a subset of the possible plots is included here to illustrate the key findings. The complete working group report, the simulation datasets, and the S-PLUS code used to summarize them are available at the BioPharmNet website[1].

Detecting DR: As illustrated in Figure 4.2, all methods were capable of controlling the Type I error rate. Fluctuations around the 5% level are consistent with Monte Carlo error.

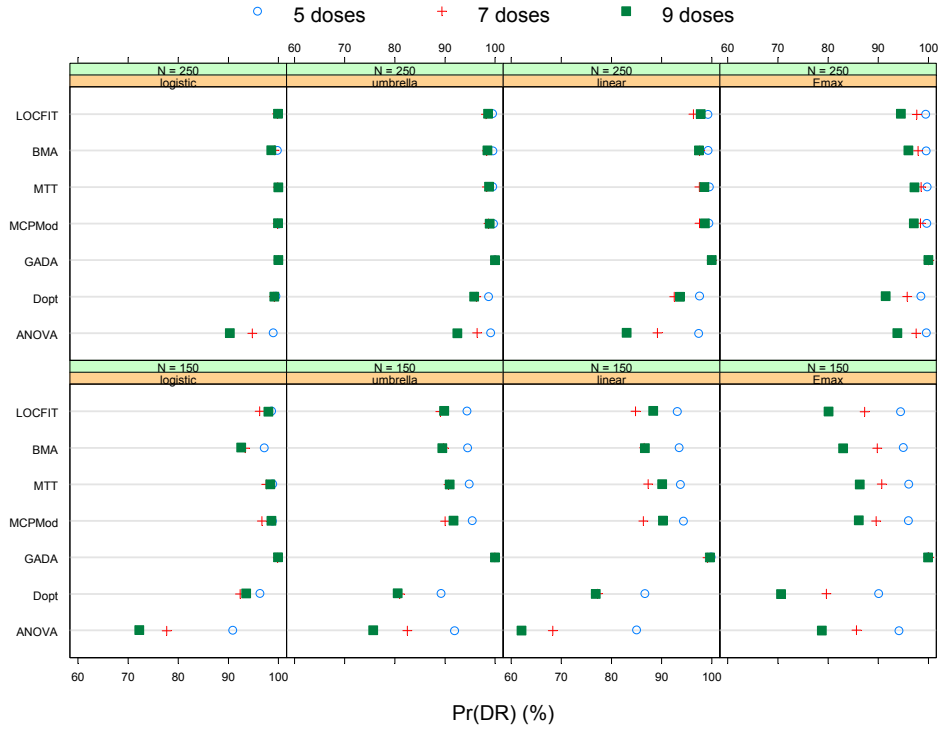
Figure 4.2. Type I error rate for detecting dose response under flat profile.



The probabilities of detecting DR under active profiles are included in Figure 4.3.

[1] http://biopharmnet.com/doc/doc12005.html

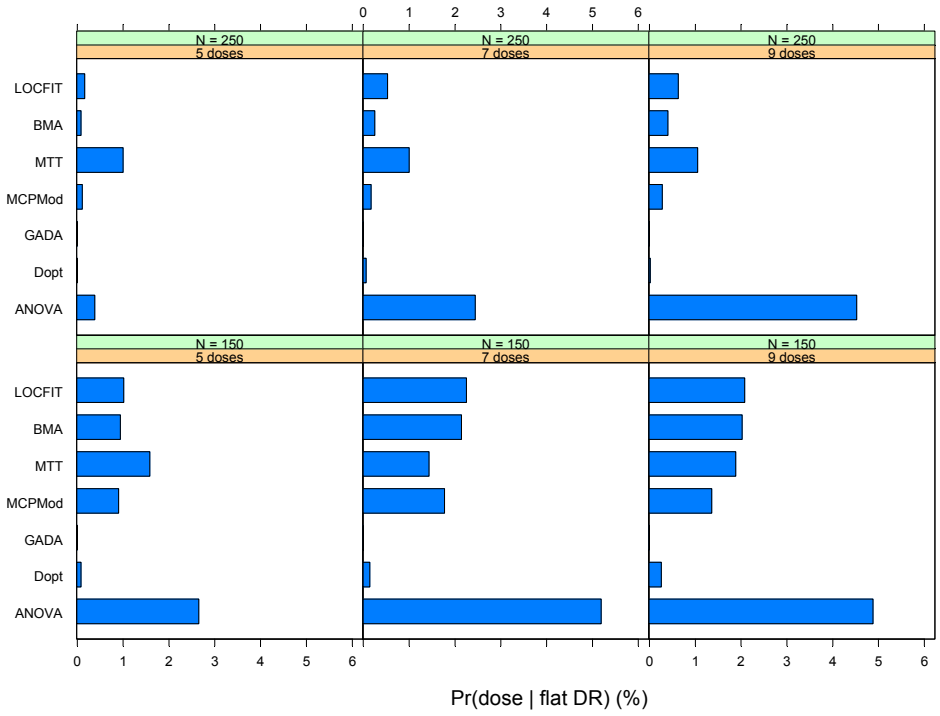
Figure 4.3. Power to detect dose response under active DR profiles.



For a total sample size of N=250 patients, all methods have reasonable power to detect dose response under the different DR models, with ANOVA presenting relative worse performance than the remaining methods. When the sample size is reduced to N=150 patients, the differences among the methods become more pronounced: the relative performance of ANOVA deteriorates, especially for larger number of doses, and Dopt shows worse relative performance under some of the DR models and with 7 or 9 doses. GADA has the best overall performance with regard to this metric. It is worth noting that increasing the number of doses, while keeping the total sample size fixed, does not lead to an increase in power to detect DR, even for model-based approaches. In fact, for some model-method combinations, *Pr(DR)* decreased with the number of doses.

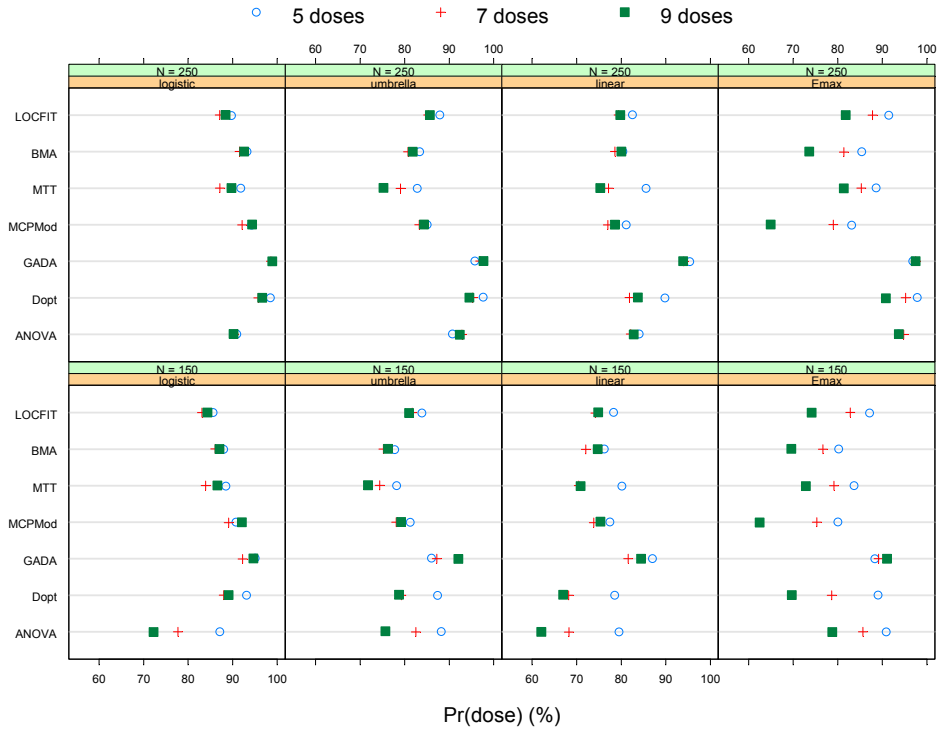
Identifying clinical relevance: Figure 4.4 shows the probabilities of incorrectly identifying a dose that produces a clinically relevant effect under a flat dose response model. The adaptive methods, GADA and Dopt, have the best overall performance with regard to this metric, having very little chance of indicating that the development program should proceed to the confirmatory stage. ANOVA has the worst performance, which deteriorates further with the increase in number of doses and reduction in N.

Figure 4.4. Probabilities of identifying clinical relevant dose under flat dose response.



The probabilities of correctly choosing a clinically relevant dose under active DR profiles are presented in Figure 4.5. The performance of the methods varies considerably with the underlying DR model, the total sample size, and the dose design. For the larger sample size, the design-focused adaptive methods (GADA and Dopt) show better overall performance, but when the sample size is reduced, only GADA is capable of maintaining it. The Emax model poses the greater amount of difficulties for all methods, but GADA. ANOVA shows probabilities of dose selection similar to analysis-focused adaptive methods, but one should recall its high false positive dose selection rates in Figure 4.4.

Figure 4.5. Probabilities of identifying clinical relevant dose under active dose response.



Selecting a target dose: The plots of the relative bias and relative absolute error in the MED estimates are displayed in Figures 4.6 and 4.7, respectively. Both plots indicate the dependence of the MED precision on the DR profile: the Emax shape leads to considerably more biased and less precise MED estimates, for all methods considered in the simulation. The design-focused adaptive methods do not show superior performance relative to the other methods for those metrics. In fact, GADA has somewhat worse performance than other methods for the Emax model, especially under the 5-dose design.

To further evaluate the performance of the methods with regard to target dose selection, we consider the correct target interval probabilities described in Section 4.2 and shown in Figure 4.8. Even though the design-focused adaptive methods clearly show better relative performance compared to the other methods, they choose a dose in the correct target interval only between 40% and 60% of the time, depending on the DR model and sample size. Among the non-adaptive methods, ANOVA has the worst overall performance, with correct target probabilities less than 30% for the logistic and linear models (0% for designs that do not include the target dose). This illustrates a fundamental problem in current DF trials, which was clearly evident in the simulations: estimating target doses is a much more difficult problem than detecting the presence of DR. Because sample size calculations for dose-ranging trials are typically based on the latter, in general they do not provide sufficient precision for estimating the MED and other target doses. We will return to this issue in Section 6. It is interesting to note that sample size did not have a major impact on the target interval probabilities, at least for the range considered in the simulations. It should be noted that the target interval probabilities shown in Figure 4.8 refer to the specific assumptions used in the simulation study and, in particular, the 10%

tolerance around the target effect. Different simulation assumptions and target effect tolerances will lead to different target interval probabilities.

Figure 4.6. Relative bias in target dose estimation.

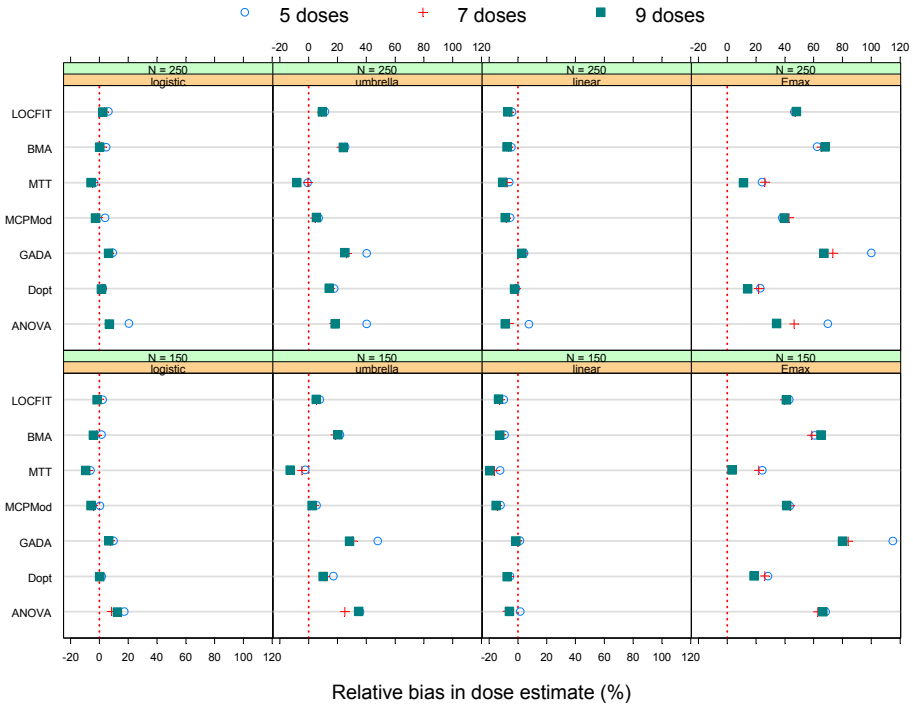


Figure 4.7. Relative absolute error in target dose estimation.

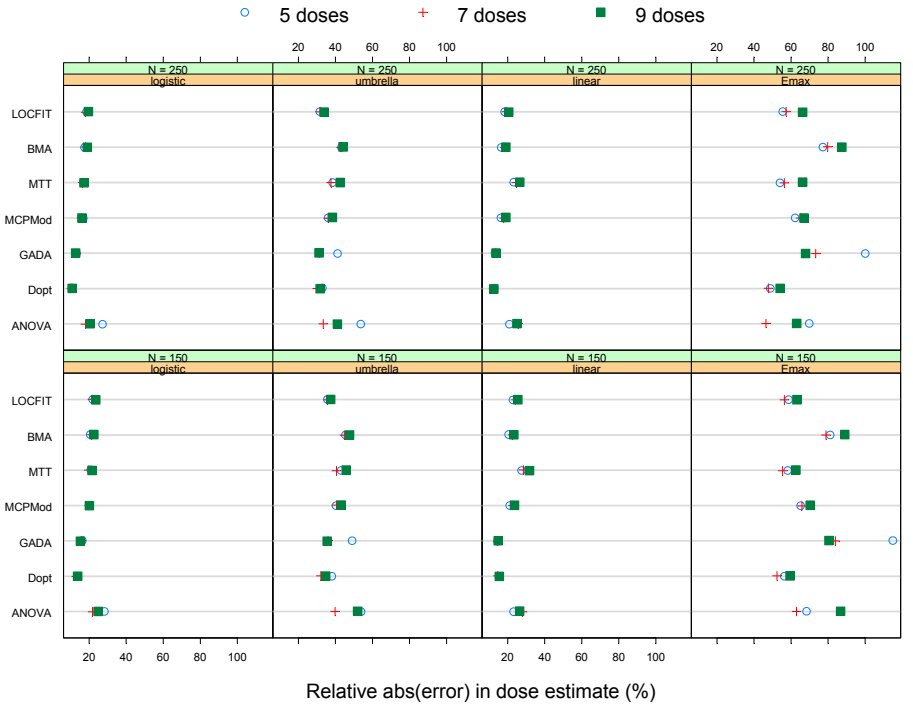


Figure 4.8. Probabilities of correct target dose interval selection.

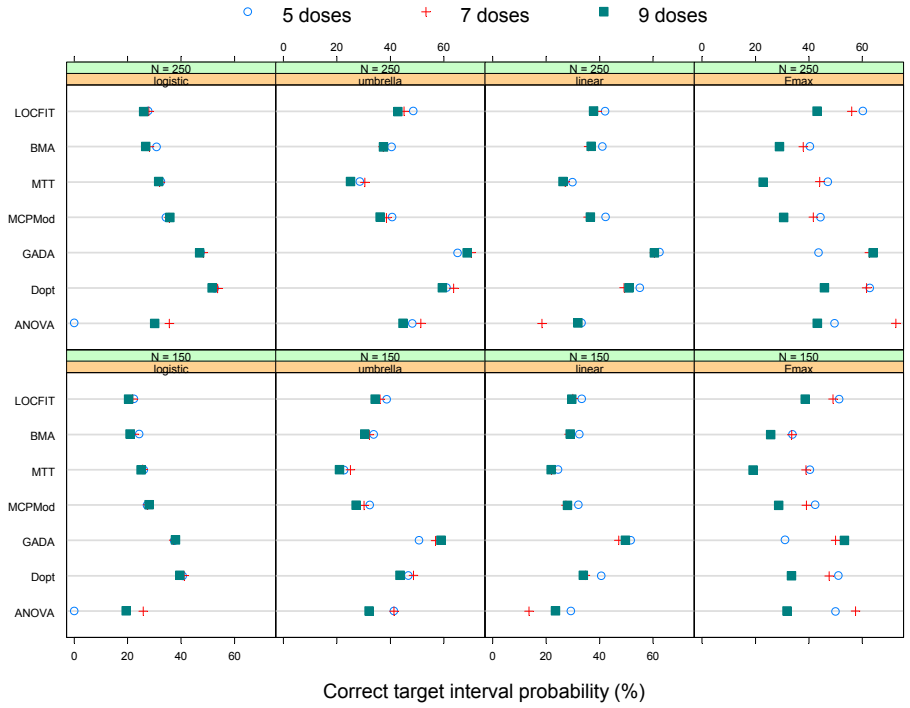
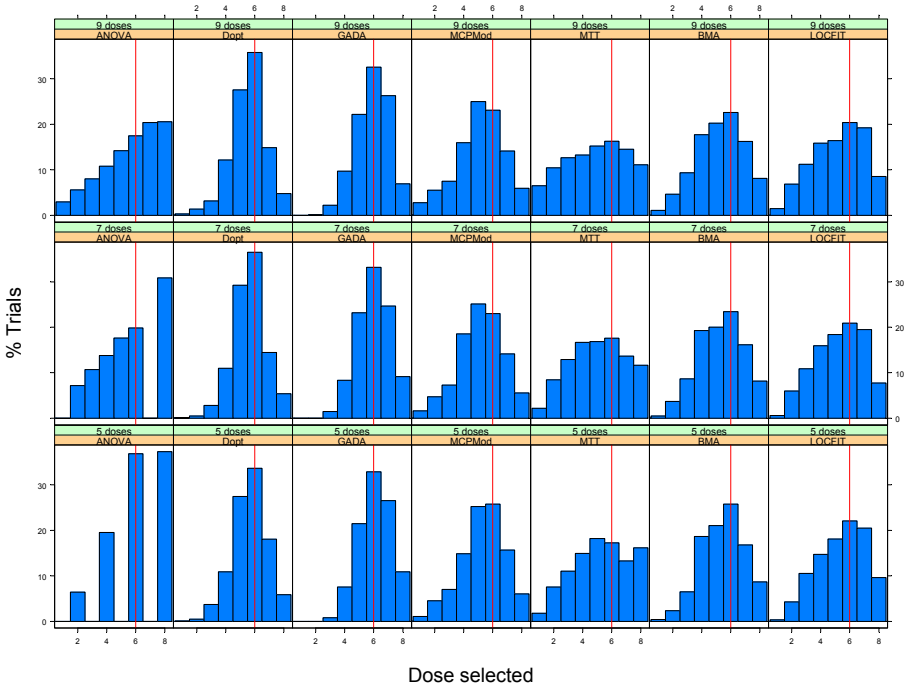


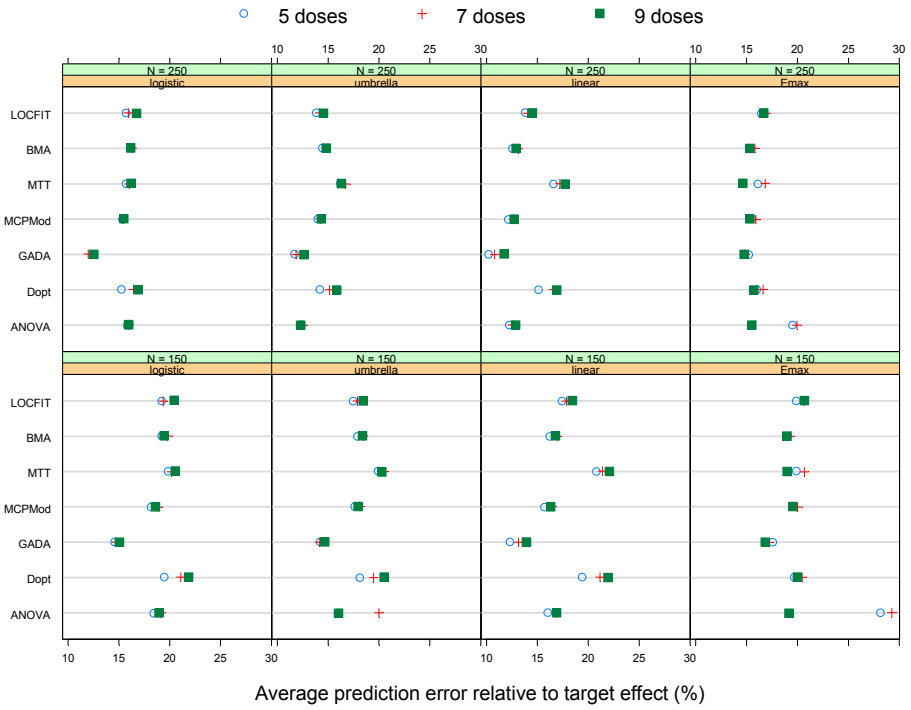
Figure 4.9. Histograms of estimated target doses for linear model, N = 150.



The histograms of the estimated target doses for the linear model (N=150), shown in Figure 4.9, provide more detailed information about the dose selection. Clearly GADA and Dopt have higher chance of selecting the correct target dose (indicated by red vertical line), but they still present considerable variability in the estimated values. The analysis-focused adaptive methods produce even more variation in the respective estimates.

Estimating the dose response: The plots of the average absolute prediction errors (pAPE) relative to the target effect, presented in Figure 4.10, suggest that there are no striking differences among the methods with regard to DR estimation. GADA shows the overall best performance, but with pAPE values not substantially smaller than MCP-Mod, for example. Somewhat surprisingly, the linear model is the one that best differentiates the performance among the methods: Dopt and MTT appear to have greater difficulty with it.

Figure 4.10. Average absolute prediction error relative to target effect.



To further illustrate the performance of the methods with regard to DR estimation, we consider prediction error quantiles plots for the logistic model with 9 doses and N=150, and the umbrella model with 5 doses and N=250. The corresponding plots are included in Figures 4.11 and 4.12. The greater precision in DR estimation associated with GADA is again illustrated in the plots: the 5%-95% quantile curves produce a narrower band for this adaptive method. The difference with respect to the other methods is not large, but certainly noticeable. There is some mild suggestion of bias in DR estimation, especially for the logistic model (with the exception of Dopt and MTT). Overall, it is clear that, given the small sample size utilized, none of the methods is capable of producing sufficiently precise estimates of the DR profile.

Figure 4.11. Prediction error quantiles for logistic model, N = 150 and 9 doses.

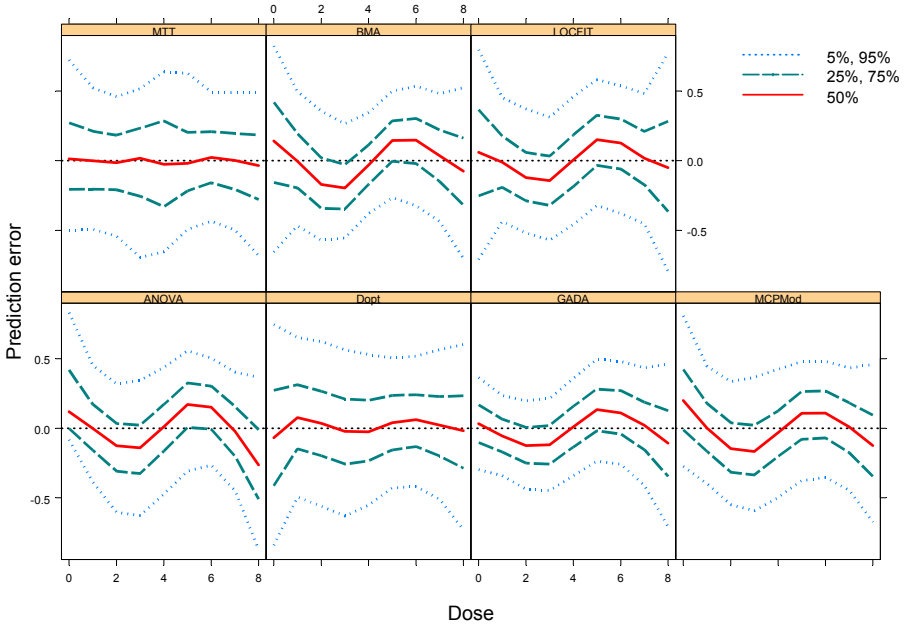
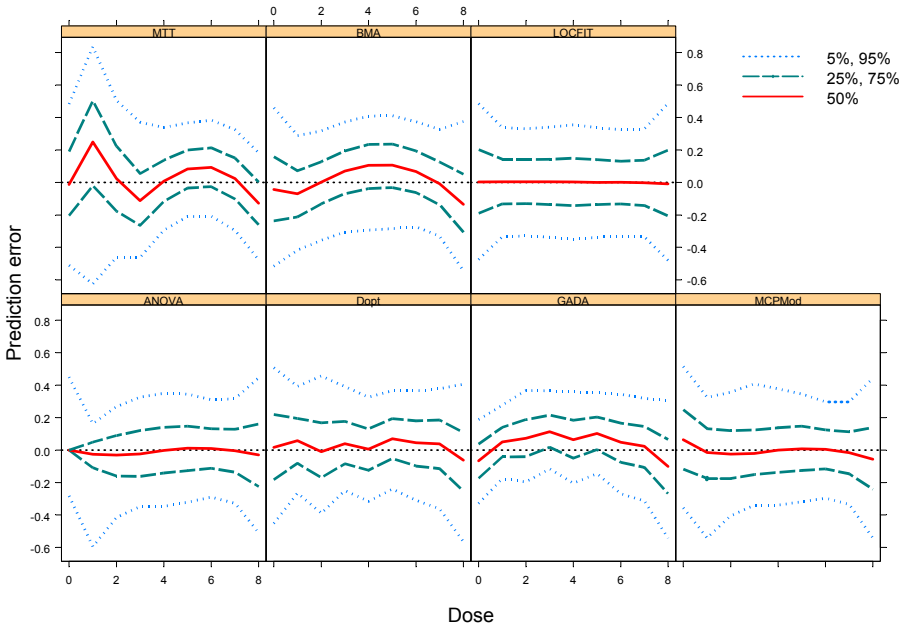


Figure 4.12. Prediction error quantiles for umbrella model, N = 250, and 5 doses.



The prediction error quantiles are calculated separately per dose and therefore do not provide information about the precision of the individual fitted curves. Because it would be impractical to present a plot of all the individual predicted DR curves produced in the simulations, we consider a random sample of 50 such curves. Figures 4.13 and 4.14 show the plots of the sampled individual predicted DR curves corresponding to the models and scenarios considered, respectively, in Figures 4.11 and 4.12.

Figure 4.13. Sample of 50 predicted curves for logistic model, N = 150 and 9 doses.

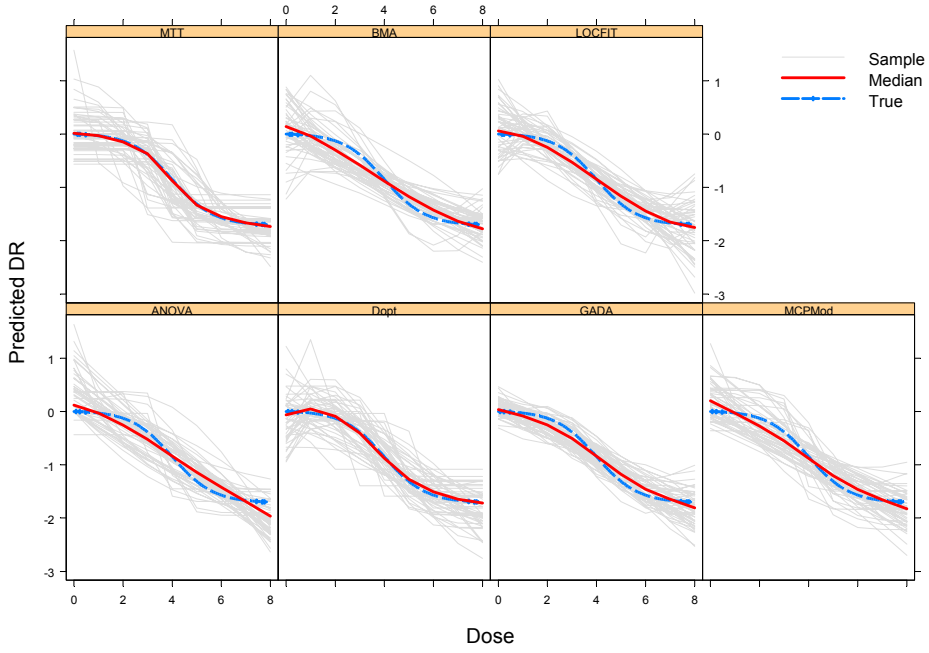
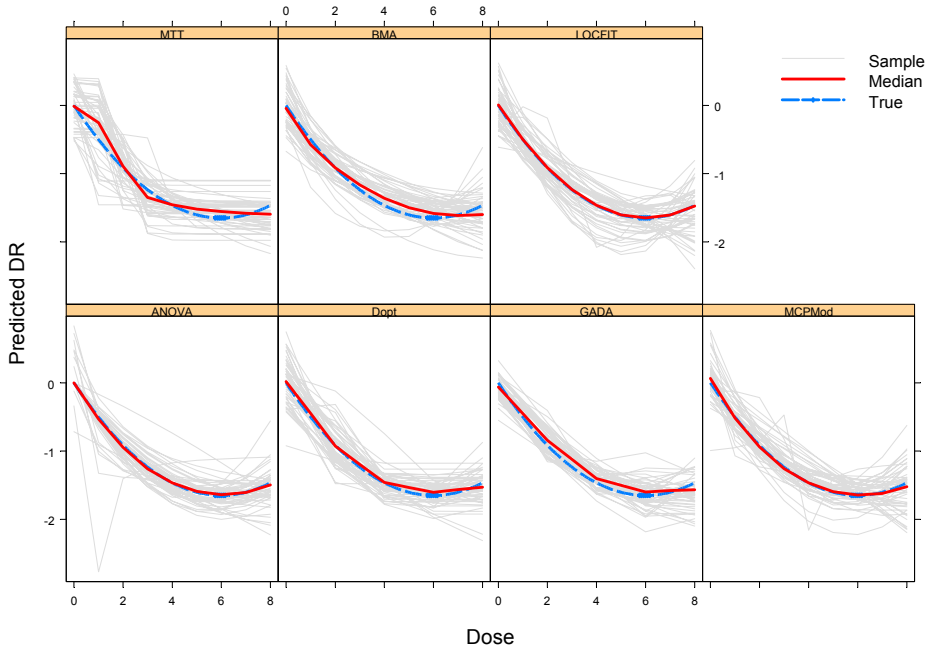


Figure 4.14. Sample of 50 predicted curves. Umbrella model, N = 250 and 5 doses.



The plots of the individual predicted DR curves reinforce the conclusions derived from the prediction quantile plots: somewhat greater precision of GADA and slight bias for some of method/model combinations. It is interesting to note the unequal precision of GADA across the dose range, a reflection of the underlying adaptive dose allocation.

4.4 Conclusions

This section summarizes the main conclusions from the simulation study. As mentioned earlier, even though the distributional assumptions and parameter values used in the simulations were motivated by a neuropathic pain dose-ranging trial, the main results and conclusions presented here transcend that particular indication. They directly apply to any parallel group dose-ranging study with a normal efficacy endpoint. Similar conclusions are expected to hold, at least qualitatively, to other types of endpoints and designs.

Table 4.2 provides an overview of the relative performance of the different methods considered in the simulations, according to the performance metrics discussed in Section 4.2. All results from the simulation study, and not just the subset presented in Section 4.3, were considered when producing the evaluations in the table. The evaluation symbols are classified into : – (poor), + (fair), ++ (good), and +++ (very good).

Table 4.2 Overview of relative performance of methods in simulation study.

Method	**Pr(DR)**	**Pr(dose)**	**Bias**	**Error**	**Dose[†] Interval**	**DR Estimation**
ANOVA	++	+	++	++	–	NA[‡]
GADA	+++	+++	++	++	++	+++
Dopt	+++	++	+++	++	++	++
MTT	+++	++	+++	++	+	++
MCP-Mod	+++	++	+++	++	+	++
BMA	+++	++	+++	++	+	++
LOCFIT	+++	++	+++	++	+	++

ANOVA, based on a Dunnett multiple comparisons adjustment, has the worst overall performance among the methods considered. The design-focused adaptive methods have the best overall performance in target dose interval selection, DR estimation and identification of clinical relevance (primarily GADA for the last two).

The first main conclusion is that *detecting dose response is considerably easier than estimating it, or identifying the target dose to bring into the confirmatory phas*e. Most methods were able to maintain good power to detect DR under the various dose response models, dose designs and sample sizes considered. However, under the designs and assumptions used in the simulations, none of them showed satisfactory performance in target dose interval selection, as well as in the estimation of the underlying DR profile.

As a consequence, the *current sample sizes used for dose-ranging studies, which are typically based on power calculations to detect the presence of DR, are inadequate for*

[†] Under the designs and assumptions used in the simulations, none of the methods had a satisfactory performance in identifying the correct dose interval for total sample sizes considered (N = 150 and 250)

[‡] ANOVA is not model-based and does not provide direct estimates for the DR profile. Modified versions were used in the simulations to allow DR estimation

dose selection and DR estimation. Larger sample sizes, and/or the selection at the end of the trial of a larger number of doses to be brought into the confirmatory phase, are needed to ensure a reasonable likelihood of bringing an appropriate dose into Phase III.

Adaptive dose-ranging designs and methods clearly lead to gains in power to detect DR and in precision to select the target dose interval and to estimate the DR, with the greatest potential for improvement being the latter two. Out of the methods considered in the simulation study, GADA had the best overall performance, especially with regard to DR estimation, the probability of DR detection, and the identification of clinical relevance. In general, model-based methods showed better performance compared to methods based on hypothesis testing.

An interesting observation in the simulation study is that, when the total sample size is kept fixed and the dose range covers the region of main dose response activity, increasing the number of doses beyond 5 does not seem to produce noticeable improvements in performance. In fact, for some of the methods (most noticeably ANOVA) performance decreased with number of doses. A possible explanation for this is the trade-off between the greater level of detail about the DR profile associated with more doses and the greater precision in the estimation of the mean DR at each dose associated with fewer doses.

5. LOGISTICAL AND REGULATORY CONSIDERATIONS

This section gives a brief discussion of implementation and regulatory issues associated with adaptive dose-ranging designs. Many of the logistical issues are shared with other adaptive designs methods, but, as discussed in Section 5.1, some are more specific to the class of adaptive designs we have considered (see Quinlan and Krams, 2006 for a more general and thorough discussion on the topic).

Because we concentrate here on designs leading into confirmatory trials, but not with confirmatory value on their own, regulatory concerns are less of an issue than would be, for example, in adaptive seamless Phase II/III designs (Gallo, 2006). However, some regulatory considerations still apply to adaptive dose-ranging designs, as discussed in Section 5.2.

5.1 Logistical considerations

The different approaches to dose-ranging studies compared in this white paper have very different demands on the logistics of planning and running a trial. Broadly, dose-ranging methodologies can be classified into 3 types:

- **Fixed**: designs with pre-determined randomization strategy and single analysis at the study end, illustrated by analysis-focused methods such as ANOVA and MCP-Mod.
- **Discrete adaptation**: randomization can be adjusted (*e.g.*, drop treatment arms, change treatment allocation ratios) at pre-planned interim analyses, represented by Dopt.
- **Continuous adaptation**: fine grain continuous adjustment of the randomization, adapting design as response data (possibly surrogates) becomes available, represented by GADA.

Planning and preparation: Fixed designs have a requirement for planning and preparation that most pharmaceutical companies are familiar with; however what this working group has found is that in the statistical planning currently almost no regard is paid to determining how well a trial will perform in identifying the correct dose. To correct this, fixed designs need additional planning and preparation, including simulation of the proposed trial design over a range of scenarios in order to characterize the design's ability to identify the correct dose to take forward to the next phase.

Unfortunately, off-the-shelf software does not currently exist to implement discrete adaptation methods in the context of dose-ranging. Usually a software package, or modules for an existing package, will be required. Discrete adaptation will need simulation of the trial design to assess fundamental characteristics of the design such as alpha, power and study size, as well as how well the correct dose is identified.

Similarly there are no standard, packaged continuous adaptation methods, and these will inevitably require a software package developing. Continuous adaptation methods will require simulation to assess the proposed trial's characteristics and to optimize the design. These simulations are likely to be orders of magnitude more compute intensive than those required for discrete adaptive or fixed designs.

Until there are packaged adaptive methods the time and cost to develop these methods at the trial planning stage are likely to be a significant hurdle to their adoption.

Randomization: Fixed designs can use pre-specified master randomization lists, with drug packs pre-allocated subject id's. Central randomization is only required where there is a desire to ensure even randomization across sub-populations of the subjects, or where drug re-supply will be necessary during the trial because of limited supplies of the drug unpredictable levels of subject recruitment at the centers.

Discrete adaptive designs can also use pre-specified master randomization lists but there are more reasons why central randomization would be preferred. If the doses being tested are packaged separately they can be pre-randomized. If the doses are made up by combining packs to save potentially wasted material (*e.g.* doses are made up by combining two tablets from tablets of three possible strengths plus placebo) then central randomization will be required. When at an interim analysis a treatment arm is dropped if using a pre-specified master randomization list, it will mean contacting the centers and withdrawing the unused treatment packs for the dropped treatment arm and the centers' behavior may not be prompt or reliable. Using a central randomization system to drop a treatment arm can be implemented immediately and correctly. Using a central randomization system also facilitates a greater range of options for adaptation such as adjusting the probabilities of randomizing to an arm rather than dropping it altogether and re-instating a previously dropped arm.

Continuous adaptive designs require central randomization systems. The randomization is continuously tuned based on the most complete response data available so the randomization must assign a subject to a dose as late as possible and must have good access to full, up to date, response data.

Data capture: Fixed designs do not require any real time data capture, only quick closure of the CRF database at the end of the trial, to allow timely analysis of the data.

Discrete adaptive designs face something of a quandary. One of the attractions of the discrete adaptive design is that it can be run manually, obviating the need for the development of a time consuming and expensive central software system. However at the times of the interim analysis it will be as necessary to have as full and up to date response data as the continuous adaptive designs require all the time. The less data is available the poorer the adaptation decision is likely to be. It is tempting to think that the now widely available electronic data capture (EDC) systems for collecting CRFs can supply this data, but EDCs are designed principally to ensure the data is validated and that the system remains validated. They were not designed for quick, incomplete, data collection during the trial, but quick completion at the end. Inevitably this tends to make them inflexible and somewhat cumbersome for our purpose. So without the support of a targeted response capture system developed specifically to support adaptation decisions, having complete data at the interims may be hard to achieve unless the trial can be restricted to a small number of well run centers, allowing the data to be gathered manually.

Continuous adaptive designs require good access to full, up to date, response data. The adaptation will only require the principal outcome for each subject however (often a single number) and not the full CRF. Often interim data is required as well as final data – to increase the data available to the system, particularly early in the trial. It is usually advantageous to collect this data separately to the CRF data even though this means some duplication. However by retaining the EDC system as the principal route for capturing the full CRF data it enables the fast response system capture to be more easily modified and extended for each trial. This in turn means it can be modified to suit the trial and be easy and convenient for investigators (and possibly subjects) to use. This in turn leads to a high level of compliance in returning the key response data required for adaptation in a timely fashion.

Adaptive dose supply: Even in fixed-design studies, if the drug supply is limited and recruitment unpredictable, an adaptive dose supply must be used to re-supply only those centers that are successfully recruiting subjects. This avoids the wastage of large initial supplies of drugs to centers that then end up recruiting few if any subjects into the trial.

Data monitoring: Fixed designs or discrete adaptive designs have the same requirements for safety data monitoring. This includes the case by case monitoring required to comply with regulatory reporting requirements, as well as the unblinded assessments required at interim analyses.

5.2 Regulatory considerations

While trial designs for early phase drug development are under the purview of the sponsoring company as long as strict compliance to regulations around potential human risk and safety is maintained, a successful dose-ranging trial brings major evidence to regulatory discussions such as end-of-phase IIA or IIB meetings. The validity of the dose recommendation for a phase III program based on adaptive strategy will be scrutinized in the following contexts: (1) whether the trial uses the same endpoint as the future phase III confirmatory trials, *e.g.*, the trial might use a biomarker or a clinical utility; (2) whether the trial itself may be considered as a confirmatory trial; (3) whether the trial will be seamlessly transitioned to a phase III confirmatory trial by either linking operationally or combining evidence from two trials.

If (2) and (3) are not proposed by the sponsor, adaptive dose-ranging designs should be considered as an effective learning paradigm for drug development where the risks of missing an accurate assessment of the true underlying dose-response profile of an investigational treatment are borne by the sponsor. These plans would not require special approval from health authorities and regulators. On the other hand if (2) or (3) is intended, then the adaptive dose-ranging trial would need to be planned in the context of the phase III program and a major discussion with regulatory authorities should be expected before the trial is initiated. The biggest validity concern is that allowing sponsors or physicians access to ongoing trial results could introduce bias that consciously or unconsciously change their conduct on an ad hoc basis. Certain logistical or statistical procedures need to be in place to remove such bias (Maca et al., 2006; Bretz et al., 2006).

In addition, it is very common for an adaptive dose-ranging trial to use an endpoint different from the endpoint used in the confirmatory trial that is required for regulatory approval. The difference is induced due to either a registration endpoint takes too long to measure, *e.g.*, survival endpoint, or a clinical utility is used to enhance development decisions based on a benefit-risk profile. In the former situation, a validated biomarker with shorter duration may be introduced for the purpose of either proof of concept or adaptive randomization of patients. Since the science and guidance development in biomarkers and clinical utility functions are still evolving, discussing these options with regulatory authorities ahead of time would be highly encouraged.

The timing of initiating these regulatory discussions is also very important. Depending on the design features, it could occur as early as a pre-IND meeting if a phase 2A/2B seamless adaptive dose-ranging trial is planned. It would trigger much earlier internal discussion on the development options and impact on the time line of IND preparation if there are large modeling and simulation work needed to be done before the regulatory discussion.

In summary, knowledge generated from an adaptive dose-ranging trial is crucial in taking the drug to approval even the trial usually occurs in the early phase of the drug development. Thus, it is important to include regulatory considerations as a part of the trial design considerations.

6. RECOMMENDATIONS

This section presents the main recommendations from the Adaptive Dose-Ranging Studies PhRMA working group. The results and conclusions of the simulation study discussed in Section 4, together with the logistical and regulatory considerations presented in Section 5, are of central importance to the group's recommendations. To recall the context in which the recommendations were derived, we reproduce below the key conclusions discussed in Section 4.4.

- Detecting dose response is considerably easier than estimating it, or identifying the target dose to bring into the confirmatory phase.
- Current sample sizes used for dose-ranging studies, which are typically based on power calculations to detect the presence of DR, are inadequate for dose selection and DR estimation.

- Adaptive dose-ranging designs and methods clearly lead to gains in power to detect DR and in precision to select target dose(s) and to estimate the DR.

As discussed earlier, even though the simulations were motivated by a specific indication (neuropathic pain), the conclusions and recommendations presented in this paper apply broadly to dose-ranging studies with similar endpoints/designs and are expected to hold, at least qualitatively, to other types of endpoints and designs. It would be useful to perform similar methods evaluations using other types of endpoints and designs. To that end, the approach described in Section 4, based on simulations under a variety of scenarios and a combination of performance metrics, can be easily generalized.

Based on the conclusions from the simulation study, as well as further internal discussions, the recommendations from the working group are as follows:

- Because they can lead to substantial gains in performance over traditional dose-ranging methods, adaptive, model-based dose-ranging approaches should be encouraged for routine use in drug development.
- Sample size calculations for Phase II studies used for dose selection should take into account the desired precision of the estimated target dose and, possibly, also the estimated DR profile. At a minimum, the precision of target dose estimate under the planned sample size should be assessed.
- When the required sample size for adequate precision in dose selection is not practically feasible, one should consider allowing two or three doses to be selected for the confirmatory phase. This would be used to ensure an appropriate probability of including the target dose among the ones selected. Adaptive designs should then be used in the confirmatory trials for greater efficiency (*e.g.*, dropping less efficacious/safe doses earlier)
- Ideally, proof-of-concept and dose selection should be combined into one seamless trial, when feasible. Development strategies that envision this should be encouraged.
- Early stopping rules, for both efficacy and futility, should be used, when feasible, to maximize the gains in efficiency associated with adaptive designs. Bayesian methods are particularly well-suited for this purpose.
- Routine use of trial simulations at the protocol design stage to determine appropriate sample sizes, to estimate the operational characteristics of designs and methods under consideration, and to evaluate their sensitivity to deviations from underlying assumptions should be encouraged. This holds true for both adaptive and non-adaptive approaches.
- Appropriate software is critical for designing, implementing, and analyzing data from adaptive dose-ranging studies needs to be developed. An effort to develop user-requirements with broad industry input for public publication should be considered.
- In practice, one should balance the potential gains associated with adaptive dose-ranging designs against their greater methodological and operational complexity.

REFERENCES

Berry DA, Mueller P, Grieve AP, Smith MK, Parke T, Krams M. (2002). Bayesian designs for dose-ranging drug trials. *Case Studies in Bayesian Statistics*. Volume 5. Gatsonis C, Kass RE, Carlin B, Carriquiry A, Gelman A, Verdinelli I, West M (editors). Springer-Verlag, New York, 99-181.

Bretz F, Pinheiro J, Branson M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*. 61, 738-748.

Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, **48**(4), 623-634.

Capizzi T, Survill TT, Heyse JF, Malani H. (1992). An empirical and simulated comparison of some tests for detecting progressive of response with increasing doses of a compound. *Biometrical Journal*. 34, 275-289.

Dunnett CW. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*. 50, 1096-1121.

Gallo P (2006). Confidentiality and trial integrity issues for adaptive designs. *Drug Information Journal*. 40(6): 445-450.

Gaydos B, Krams M, Perevozskaya I, Bretz F, Liu Q, Gallo P, Berry D, Chuang-Stein C, Pinheiro J, Bedding A (2006). Adaptive Dose-Response Studies. *Drug Information Journal*. 40(6): 451-461.

Krams M, Lees KR, Hacke W, Grieve AP, Orgogozo JM, Ford GA for the ASTIN Study investigators. (2003). ASTIN: An adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke*. 34, 2543-254.

Loader C. (1999). *Local Regression and Likelihood*. Springer, New York.

Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., and Krams, M. (2006) Adaptive Seamless Phase II/III Designs – Background, Operational Aspects, and Examples. *Drug Information Journal*. 2006; 40(4): 463 – 473

Quinlan JA and Krams, M (2006). Implementing adaptive designs: logistical and operational considerations. *Drug Information Journal*. 40(6): 437-444.

Stewart WH, Ruberg SJ. (2000). Detecting dose response with contrasts. *Statistics in Medicine*. 19, 913-921.

Tukey JW, Ciminera JL, Heyse JF. (1985). Testing the statistical certainty of a response to increasing dose of a drug. *Biometrics*. 41, 295-301.

West M, Harrison PJ. (1997). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.

APPENDIX: Acronyms used in the paper

ANOVA	Analysis of Variance
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BMA	Bayesian Model-Averaging
CRF	Case Report Form
Dopt	D-optimal response adaptive
DR	Dose response
EDC	Electronic Data Capture
FWER	Family-Wise Error Rate
GADA	General Adaptive Dose Allocation
IND	Investigational New Drug
LOCFIT	Local Regression Fit
LOESS	Local Regression
MCMC	Markov Chain – Monte Carlo
MCP	Multiple Comparison Procedure
MCP-Mod	Multiple Comparison Procedure – Modeling
NDLM	Normal Dynamic Linear Model
MED	Minimum Efficacious Dose
MTT	Multiple Trend Test
PhRMA	Pharmaceutical Research and Manufacturers of America
PISC	Pharmaceutical Innovation Steering Committee
PoC	Proof-of-Concept
VAS	Visual Analog Scale