Dose Finding – A Challenge in Statistics

Frank Bretz[*,1], **Jason Hsu**[2], **José Pinheiro**[3], and **Yi Liu**[2]

[1] Clinical Information Sciences, Novartis Pharma AG, CH-4002 Basel, Switzerland
[2] Department of Statistics, Ohio State University, Columbus, OH 43210, USA
[3] Clinical Information Sciences, Novartis Pharmaceuticals, East Hanover, New Jersey, NJ 07936 USA

Received 29 February 2008, revised 18 May 2008, accepted 21 May 2008

Summary

A good understanding and characterization of the dose response relationship of any new compound is an important and ubiquitous problem in many areas of scientific investigation. This is especially true in the context of pharmaceutical drug development, where it is mandatory to launch safe drugs which demonstrate a clinically relevant effect. Selecting a dose too high may result in unacceptable safety problems, while selecting a dose too low may lead to ineffective drugs. Dose finding studies thus play a key role in any drug development program and are often the gate-keeper for large confirmatory studies.

In this overview paper we focus on definitive and confirmatory dose finding studies in Phase II or III, reviewing relevant statistical design and analysis methods. In particular, we describe multiple comparison procedures, modeling approaches, and hybrid methods combining the advantages of both. An outlook to adaptive dose finding methods is also given. We use a real data example to illustrate the methods, together with a brief overview of relevant software.

Key words: Adaptive designs; Adaptive dose finding; Clinical trial; Dose ranging; Dose response; Drug development; Minimum effective dose; Modeling; Multiple comparisons; Proof-of-concept; Randomized trial.

1 Introduction

A good understanding and characterization of the dose response relationship is a fundamental step in the investigation of any new compound, be it a medicinal drug, an herbicide or fertilizer, a molecular entity, an environmental toxin, or an industrial chemical. Already 500 years ago, the humanist, physician and chemist Paracelsus (Figure 1) was well aware of the basic principle that any compound, even the apparently most harmless, is potentially toxic if administered at high enough doses, as illustrated by a well-known quote usually attributed to Paracelsus (Letter, 2000, p. 162):

> All things are poison and nothing is without poison,
> only the dose permits something not to be poison.

That is to say, compounds typically considered as toxic can be harmless, and potentially even beneficial, if administered at small doses. Conversely, compounds typically considered to be non-toxic, or even essential for living, can be lethal at sufficiently high doses. Examples of apparently harmless substances are water and table salt, which become lethal for humans at doses of $10\,\ell$ and $300\,g$, respectively (Unkelbach and Wolf, 1985).

The history of penicillin illustrates the importance of dose finding from a different perspective. Although the discovery of penicillin is credited to Sir Alexander Fleming in 1928, its use in humans (and hence its usefulness) was only made possible after 1939, when a team of researchers at Oxford

[*] Corresponding author: e-mail: frank.bretz@novartis.com, Phone: ++41 61 3244064, Fax: ++41 61 324 3039

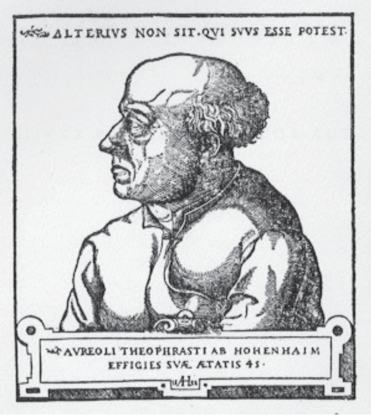
ALTERIVS NON SIT.QVI SVVS ESSE POTEST.
AVREOLI THEOPHRASTI AB HOHENHAIM
EFFIGIES SVÆ ÆTATIS 45.

Figure 1 Portrait of Paracelsus (1493–1541). Source: Wikipedia.

University, led by Howard Florey and Ernst Chain, identified the correct dose for *in vivo* application. In recognition of their importance to the development of penicillin, Florey and Chain shared the 1945 Nobel prize in medicine with Fleming (Radetsky, 1996).

Finding the right dose is therefore an ubiquitous and important problem, even more so in the context of drug development, when patients will be exposed to a medicinal drug once it has been released on the market. Determining an adequate dose level for a drug and, more broadly, characterizing its dose response relationship with respect to both efficacy and safety, are thus key objectives of clinical drug development. If the dose is set too high, safety and tolerability problems are likely to result, while selecting too low a dose makes it difficult to establish adequate efficacy in the confirmatory phase, possibly leading to a failed development program. It is well known that many potentially efficacious drugs have been discarded, and much time and resources have been wasted by pharmaceutical companies, because of incorrect dosing. In addition, many marketed drugs needed to have their label changed due to inappropriate, generally excessive, initial dose recommendations. This has been documented by the U.S. Food and Drug Administration (FDA), who reported that approximately 10% of drugs approved between 1980–1989 have undergone dose changes – mostly decreases – of greater than 33% (FDC Report, 1991).

An indication of the importance of properly conducted dose response studies is the early publication of the ICH-E4 guideline (ICH-E4, 1994), which is the primary source of regulatory guidance in this area. The guideline gets very specific already in the introduction when it motivates the importance of dose response information:

> Historically, drugs have often been initially marketed at what were later recognized as excessive doses . . . This situation has been improved by attempts to find the smallest dose with a discernible useful effect or a maximum dose beyond which no further beneficial effects is seen . . .

It becomes transparent from this quote, and the remainder of the ICH-E4 guideline, that regulatory agencies recognize the need to obtain appropriate dose response information as a critical part of clinical drug development. But even if it is generally agreed that understanding the relationships among administered dose, drug-concentration in blood, and clinical response is important, the objective setting for an actual trial may be subject to much debate. Ruberg (1995a), for example, considered the following questions to be relevant in the context of dose finding:

(i) Is there any evidence of a drug effect?
The detection of a dose response signal is often related to the determination of proof-of-concept (PoC) in a development program. This is a critical decision point, since a positive PoC

coupled with a subsequent commitment to go into full development lead to substantial financial investments.

(ii) *What doses are (relevantly) different from control?*
This question is closely connected to the estimation of a minimum effective dose, that is, "the smallest dose with a discernible useful effect" (ICH-E4, 1994). If confirmatory pairwise comparisons with a control are of main interest (such as in Phase III trials), multiple comparison procedures may be appropriate to answer this question.

(iii) *What is the dose response relationship?*
This question is broader than the previous one in the sense that it asks for a complete functional description of the dose response relationship. If this is of main interest, modeling approaches may be appropriate to take full advantage of the observed data.

(iv) *What is the optimal dose?*
Although very natural, this question is likely to be the most difficult to answer. In practice this question may not even be well defined in the sense that different groups of people may have a different understanding of what "optimal dose" means. In all circumstance, any answer to this question will be a trade-off between efficacy considerations, safety issues and regimen convenience.

Consequently, dose finding studies can be encountered at almost every stage of drug development, such as pre-clinical toxicology studies in animals at different doses, dose-escalation studies in healthy volunteers to estimate the maximum tolerated dose in Phase I trials, special Phase I designs required due to unmet medical needs as encountered in oncology, exposure response studies accounting for pharmacokinetic data, and dose finding studies typically encountered in the late stage of drug development, that is, in Phase II and/or Phase III clinical trials. Given this variety of applications within a drug development program, the complexity induced by multiple study objectives and the critical importance due to subsequent substantial financial commitments, it becomes evident that dose finding studies pose a serious challenge for a proper design and analysis. In fact, both the FDA and the Pharmaceutical Research and Manufacturers of America (PhRMA) have identified poor dose selection resulting from incorrect or incomplete knowledge of the dose response relationship, for both efficacy and safety, as one of the key drivers of the high attrition rates currently plaguing late phase clinical trials across the pharmaceutical industry (FDA, 2004; Bornkamp et al., 2007). In this regard, FDA released a White Paper entitled "Stagnation/Innovation: Challenge and Opportunity on the Critical Path to New Medical Products" (FDA, 2004). The document acknowledges that today's revolution in biomedical science has raised new hope for the treatment of many diseases, but points out that the number of new drug and biologic applications submitted to the FDA has declined considerably in the last decade and discusses several potential causes for this decline. The White Paper concludes that if the drug development processes will not become more efficient and effective, innovation may continue to stagnate and the biomedical revolution may fail to achieve its full potential. This White Paper has triggered many activities across the pharmaceutical industry to enhance current dose finding practices. For example, PhRMA has constituted a working group on "Adaptive Dose Ranging Studies" to evaluate and propose recommendations to address the problem of inadequate dose response information. The working group investigated different types of adaptive dose ranging designs and methods focusing on Phase II trials by means of an extensive simulation study evaluating the performance of the methods under a variety of trial scenarios. One major conclusion from the PhRMA working group activities is that innovative dose finding methods typically outperform traditional methods, especially with respect to the precision of estimating the functional dose response relationship and related target doses (Bornkamp et al., 2007).

In light of these ongoing discussions and activities, this article aims at reviewing some of the key methodologies used in dose finding trials typically encountered in the late stage of drug development. We will discuss in detail different methodologies which could be applied to either Phase II and/or Phase III studies. This includes multiple comparisons procedures, modeling techniques, hybrid approaches combining multiple comparisons with modeling, and response-adaptive dose finding studies.

Out of scope for this overview are dose finding studies in early development, which often take place under different constraints and use different methodologies, such as the traditional 3 + 3 designs, up-and-down designs or continual reassessment methods. For general reading about dose finding in drug development we refer to the edited books by Ting (2006), Chevret (2006), and Krishna (2006) and the references therein.

Accordingly, this paper is organized as follows. In Section 2 we introduce a dose finding study, which will be used later to illustrate some of the concepts and results. In Sections 3 and 4 we will review two major approaches of analyzing dose response studies. In Section 3 we discuss multiple comparison procedures applied to dose response testing and dose finding. Section 4 is devoted to modeling approaches. In Section 5 we discuss recently introduced hybrid dose finding methods that combine principles of multiple comparisons with modeling techniques. Section 6 is devoted to response-adaptive dose finding studies, which extend some of the previously discussed methods. Concluding remarks are given in Section 7.

2 A Dose Finding Study

In this section we introduce a dose finding study to motivate the subsequent concepts and discussions. We come back to this example in the following sections as needed. Consider the dose response data in Table 1 taken from Ruberg (1995b). The data are from a pre-clinical dose finding study comparing nine active dose groups with a zero-dose control (vehicle group). Six animals were investigated for each of the 10 dose groups.

Since the original data were not published, we digitalized the dose response plot from Ruberg (1995b) and fine tuned the resulting individual observations to obtain response values matching the summary statistics reproduced in Table 1. A plot of the new generated data is given in Figure 2. The individual data are available from http://www.biostat.uni-hannover.de/staff/bretz/data.htm.

We use this plot to illustrate the potential study objectives related to dose finding studies, as introduced in Section 1. The discussion below serves to motivate some of the considerations in the subsequent sections. First, it could be of interest to assess a global dose response signal. This is typically achieved by using hypotheses testing approaches, since the trade-off between false positive and false negative decisions based on incorrectly declaring or missing dose response needs to be carefully controlled. For the example in Figure 2, a dose response signal is clearly evident, that is, we can safely assume that the factor "dose" has a significant influence on the response. Second, if a dose response signal has been shown, one might naturally be interested in identifying those doses, which provide a response different from control. From Figure 2 one would conclude that, for example, any dose higher than or equal to 3 mg/kg gives a response which is relevantly different from the zero-dose group. Proper statistical analysis methods can be applied to give a more precise answer than simply looking at the plot, with the possibility of including clinical relevance considerations as well. Either multiple comparison procedures or modeling techniques can be used to answer this question, depending on, for example, whether the study is intended to be confirmatory or not. Third, the nature of the dose response relationship is often of interest by itself. For illustration purposes, we have included a logistic model fit in Figure 2. Whether or not a particular dose response fit corresponds to the true underlying functional model is a difficult assessment, which leads to often underestimated model selection problems. Fourth, the dose with the optimal benefit/risk ratio has to be identified, which is then likely

Table 1 A pre-clinical dose finding study.

Dosage (mg/kg)	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Mean	25.5	23.9	27.7	33.4	40.5	57.9	74.4	73.4	73.5	76.2
Standard Deviation	2.6	4.0	3.3	2.3	10.5	9.9	14.6	7.6	4.5	7.9

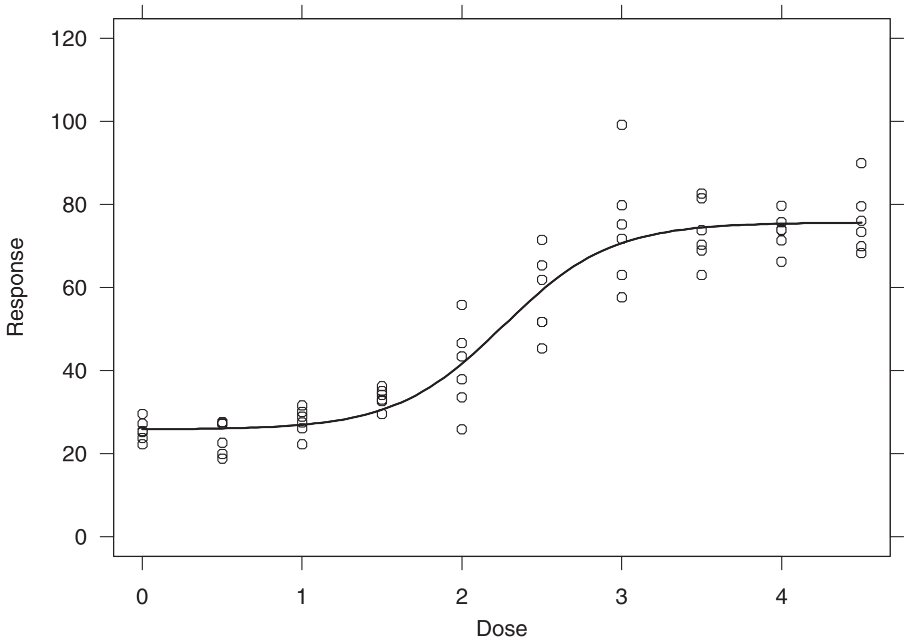

Figure 2 A pre-clinical dose finding study including a logistic model fit to the individual observations.

to be carried forward to the next development stage, such as the Phase III program, for example. As explained in Section 1, “optimality” is difficult to formalize and often the decision with which dose to continue remains a compromise accounting for scientific and non-scientific justifications.

3 Multiple Comparison Procedures

The analysis of dose finding studies can be classified into two major strategies: modeling techniques (Pinheiro et al., 2006a; Bates and Watts, 1988) and multiple comparison procedures (MCP) (Hochberg and Tamhane, 1987; Hsu, 1996). Modeling techniques take the dose as a quantitative factor. They often assume a functional relationship between “dose” and “response”, according to a pre-specified parametric model (typically to be defined in the study protocol), which can then be used to answer some of the questions from Section 1. We will review modeling techniques with more details in Section 4. In this section we consider MCP in analysis-of-variance (ANOVA) settings, which regard the dose as a qualitative factor and make very few, if any, assumptions about the underlying dose response model. MCP can be used for detecting an overall dose related signal as well as for estimating target doses of interest. Stepwise testing strategies can be applied which preserve the overall type I error rate at a pre-specified level α. Such procedures are relatively robust to the underlying dose response shape, but they are not designed for extrapolation of information beyond the observed dose levels. Inference is thus confined to the selection of the target dose among the dose levels under investigation.

3.1 Parallel designs for dose response studies

A *parallel* design is one in which subjects are assigned to one of two or more arms, each arm being allocated a different treatment (ICH-E9, 1998). To eliminate bias, subjects in each treatment group should be randomly drawn from the available subjects. If we select the subjects to be assigned to each treatment arm by simple random sampling, then statisticians call this a *completely randomized* design.

A simple random sample of size n_i consists of n_i individuals from the population chosen in such a way that every set of n_i individuals has an equal chance to be the sample actually selected. We select a simple random sample by labeling all the available subjects and using a table of random digits to select (without replacement) a sample of the desired size. Simple random sampling not only eliminates bias, but also helps to ensure that the model

$$Y_{ir} = \mu_i + \varepsilon_{ir}, \quad i = 1, 2, \ldots, k, \quad r = 1, \ldots, n_i, \tag{1}$$

is reasonable, provided the population size is infinite (or at least the sample size is small relative to the population size). Here, ε_{ir}, $i = 1, \ldots, k$, $r = 1, \ldots, n_i$, denote independent and identically distributed random variables, which, in Section 3.2, are assumed to be normally distributed.

3.2 ANOVA approach

Recall from Section 1 that the lowest dose at which a drug is effective is called the minimum effective dose (MED). For notation, let $\mu^e_{\text{negative}}, \mu^e_{\text{active}}, \mu^e_i$ be the mean efficacy responses of the negative control, active control, and the *i*th dose of the new drug, respectively.

In the absence of a treatment proven to be efficacious, if a higher treatment response is better, then the MED may be defined relative to a negative control (a placebo), as the lowest dose *i* so that, for all doses *j* greater than or equal to *i*, $\mu^e_j > \mu^e_{\text{negative}} + \Delta$ where Δ is a non-negative quantity defining what constitutes a clinically relevant difference. (If a lower treatment response is better, then the MED may be defined relative to a negative control (a placebo), as the lowest dose *i* so that, for all doses *j* greater than or equal to *i*, $\mu^e_j < \mu^e_{\text{negative}} - \Delta$ where Δ is a non-negative quantity defining what constitutes a clinically relevant difference.)

When a proven treatment (an active control) exists, then in a so-called superiority trial for which a higher treatment response is better, MED is defined as the lowest dose *i* so that, for all doses *j* greater than or equal to *i*, $\mu^e_j > \mu^e_{\text{active}} + \Delta$ where Δ is a non-negative quantity. Alternatively, in a so-called non-inferiority trial for which a higher treatment response is better, MED is defined as the lowest dose *i* so that, for all doses *j* greater than or equal to *i*, $\mu^e_j > \mu^e_{\text{active}} + \Delta$ where Δ can be a negative quantity. (In a superiority trial for which a lower treatment response is better, MED is defined as the lowest dose *i* so that, for all doses *j* greater than or equal to *i*, $\mu^e_j < \mu^e_{\text{active}} - \Delta$ where Δ is a non-negative quantity. In a so-called non-inferiority trial for which a lower treatment response is better, MED is defined as the lowest dose *i* so that, for all doses *j* greater than or equal to *i*, $\mu^e_j < \mu^e_{\text{active}} - \Delta$ where Δ can be a negative quantity.)

In the discussion below, μ_1 denotes μ^e_{negative} in a trial with a placebo control. It denotes μ^e_{active} in a trial with an active control. Finally, we note that the definition of the MED used here is not unique. When describing modeling approaches in Section 4, we will use an alternative formulation, see defintion (8).

3.2.1 Dunnett's method

Dunnett's (1955) method uses the two-sample *t* method to compare each dose group with the control group, adjusting for multiplicity. One can conceptually think of Dunnett's method simply as a bunch of two-sample inferences, adjusted for multiplicity. We will present the confidence set version of Dunnett's method, from which one can deduce a multiple testing version.

Assuming dose group 1 is the control group to be compared against, Dunnett's (1955) methods provide $100(1-\alpha)\%$ simultaneous confidence bounds or intervals on $\mu_i - \mu_1$, $i = 2, \ldots, k$, depending on whether the desired inference is 1-sided or 2-sided.

If a larger response is better, then Dunnett's 1-sided method gives the following simultaneous confidence lower bounds for the difference between each dose mean μ_i and the control mean μ_1:

$$\mu_i - \mu_1 > \hat{\mu}_i - \hat{\mu}_1 - c\hat{\sigma}\sqrt{n_i^{-1} + n_1^{-1}} \quad \text{for} \quad i = 2, \ldots, k, \tag{2}$$

where the critical value c adjusts for the multiplicity of $k - 1$ simultaneous inferences. One then infers dose levels with lower confidence bounds greater than Δ to be efficacious. If one infers the *lowest* dose i such that Dunnett's lower confidence bounds for $\mu_j - \mu_1$, $i \le j \le k$, are greater than Δ to be MED^U, then it is a conservative $100(1 - \alpha)\%$ upper confidence bound on the true MED. That is, one has $100(1 - \alpha)\%$ confidence that MED^U is a higher dose than the true MED.

We now discuss several options for the critical value c. Without multiplicity adjustment, c would simply be the univariate t upper α quantile $t_{1-\alpha,\nu}$. Applying the Bonferroni inequality to adjust for multiplicity is guaranteed to result in conservative inference. This amounts to using the quantile $t_{1-\alpha/(k-1),\nu}$ in place of c and pretending the events

$$E_i = \{\text{Inference on } \mu_i - \mu_1 \text{ is incorrect}\}, \quad i = 2, \ldots, k,$$

are *disjoint*, which they are not because it is certainly possible for inferences on two or more $\mu_i - \mu_1$ to be wrong at the same time. The Bonferroni approximation is generally not very accurate for multiple comparisons with a control. Applying Šidák's inequality to adjust for multiplicity is also guaranteed to result in conservative inference. This amounts to using the quantile $t_{(1-\alpha)^{1/(k-1)},\nu}$ and pretending the events

$$C_i = E_i^c = \{\text{Inference on } \mu_i - \mu_1 \text{ is correct}\}, \quad i = 2, \ldots, k,$$

are *independent*, which they are not because the inferences share common estimates for μ_1 and σ^2. The independence approximation is slightly better than the Bonferroni approximation, but generally still not very accurate for multiple comparisons with a control.

The critical value c that adjusts for multiplicity exactly, so that the simultaneous confidence bounds (2) have coverage probability exactly $100(1 - \alpha)\%$, depends on α, the correlation matrix of $\hat{\mu}_i - \hat{\mu}_1$, $i = 2, \ldots, k$, and the error degrees of freedom ν. If the correlation matrix has a 1-factor structure in the sense that it can be written in the form of $\lambda\lambda' + \Omega$, where Ω is a diagonal matrix and λ is a column vector, $\lambda = (\lambda_2, \ldots, \lambda_k)'$, then $c = c_{\lambda,\alpha,\nu}$ is the solution to the equation

$$\int_0^{\infty} \int_{-\infty}^{+\infty} \prod_{i=1}^{k-1} [\Phi((\lambda_i z + cs)/(1 - \lambda_i^2)^{1/2})] \, c\Phi(z) \, \gamma(s) \, \mathrm{d}s = 1 - \alpha \,. \tag{3}$$

Since the critical value c depends, in general, on $k + 2$ arguments, it is hard to tabulate c comprehensively. Standard software packages can be used instead to solve for c, such as the `mvtnorm` package in R (Hothorn et al., 2001) or the `ProbMC` function in SAS. The syntax of the `ProbMC` function, for example, is

```
probmc(distribution, q, prob, df, nparms<, parameters>)
```

The argument `distribution` should be "`dunnett1`" for one-sided Dunnett's method; the quantile `q`, is what we want to compute and is not specified; `prob` is the cumulative probability, which is $1 - \alpha$; the degrees of freedom `df` is the error degrees of freedom. In the case of a one-way ANOVA model, the correlation matrix of $\hat{\mu}_i - \hat{\mu}_1$, $i = 2, \ldots, k$, does have a 1-factor structure, with

$$\lambda_i = \left(1 + \frac{n_1}{n_i}\right)^{-1/2}, \quad i = 2, \ldots, k \,. \tag{4}$$

Even though it is impossible to tabulate c in advance in general, a computer program can solve for c in (3) at execution time depending on the sample size pattern $\boldsymbol{n} = (n_1, \ldots, n_k)$ of the data to be analyzed.

To compute lower Dunnett confidence bounds from data using SAS, assuming the doses are labeled $1, \ldots, k$ with dose 1 being the control or placebo, the syntax is

```
PROC GLM;
CLASS dose;
MODEL response = dose;
LSMEANS dose/PDIFF=CONTROLU('1') ADJUST=DUNNETT CL;
```

Simultaneous confidence upper bounds for Dunnett's 1-sided method can be computed in a similar way. Note that the use of the `MEANS` statement in `PROC GLM` is discouraged to avoid potential misleading multiple comparison results (Hsu, 1996).

Alternatively, simultaneous confidence lower or upper bounds for Dunnett's 1-sided method can be computed with the *multcomp* package implemented in R (Bretz et al., 2008; Hothorn et al., 2008). Assume that *data* is a data frame containing the response variable *resp* and the factor *dose*. The *glht* function from *multcomp* takes a fitted response model (based on *data*) and performs Dunnett's 1-sided method with the syntax

```
lm.fit <- lm(resp ~ dose, data)
lm.dunnett <- glht(lm.fit, linfct = mcp(dose = "Dunnett"), alternative = "greater")
confint(lm.dunnett)
```

3.2.2 Dose finding study example (continued)

We revisit the dose finding example from Section 2. Assuming a clinically meaningful difference of $\Delta = 40$, dose i can be confidently inferred to be superior if a *lower* confidence bound on $\mu_i - \mu_1$ is greater than 40. Applying the `PROC GLM` introduced above, Dunnett's 95% lower confidence bounds on $\mu_i - \mu_1$, $i = 2, \ldots, 10$, are as follows:

```
   i     j      Difference         Simultaneous 95%
                Between            Confidence Limits for
                Means              LSMean(i) - LSMean(j)

   2     1     -1.666707         -12.799756         Infinity
   3     1      2.151515          -8.981534         Infinity
   4     1      7.856362          -3.276687         Infinity
   5     1     14.974230           3.841181         Infinity
   6     1     32.393775          21.260726         Infinity
   7     1     48.884348          37.751299         Infinity
   8     1     47.834935          36.701886         Infinity
   9     1     47.927485          36.794436         Infinity
  10     1     50.667818          39.534769         Infinity
```

Since none of the lower confidence bounds are greater than 40, no dose can be inferred to be superior to the zero-dose group. Note that the same results are obtained when using the *multcomp* package implemented in R.

3.2.3 A stepdown method with pre-determined steps

It is possible for Dunnett's method to infer a dis-contiguous set of doses to be efficacious due to sampling variation. For example, Dunnett's method might infer doses 2, 3, and 5 to be superior to the control. How to deduce a therapeutic window in such a situation is problematic. There is thus some advantage to using a statistical method designed to give a contiguous set of doses as efficacious, such as the method described below as well as methods based on modeling approaches described in Section 4.

In this section, we describe a stepwise method designed for MED^U inference. As the name minimum effective dose implies, to infer dose MED^U to be the MED is to infer doses $\mathrm{MED}^U, \ldots, k$ to be efficacious. The set of doses inferred to be efficacious by this stepwise method is naturally contiguous. For notational convenience, $\mathrm{MED}^U = k + 1$ refers to the situation where no dose is inferred to be efficacious by the stepwise method.

Suppose dose i is considered effective if $\mu_i > \mu_1 + \Delta$. To logically infer dose k is effective by the rejection of a null hypothesis, the null hypothesis tested has to be H_{0k}: Dose k is ineffective. To logically infer doses k and $k-1$ are effective by the rejection of the null hypotheses H_{0k} and $H_{0(k-1)}$, the union of the null hypotheses H_{0k} and $H_{0(k-1)}$ needs to include the possibilities dose k is ineffective and/or dose $k-1$ is ineffective, and so forth. This is an example of forming null hypotheses using the partitioning principle of Stefansson, Kim and Hsu (1988), and Finner and Strassburger (2002).

Consider testing the null hypotheses

$H_{0k}^{\downarrow}$: Dose k is ineffective

$H_{0(k-1)}^{\downarrow}$: Dose k is effective but dose $k-1$ is ineffective

$\vdots$

$H_{0i}^{\downarrow}$: Doses $i+1, \ldots, k$ are effective but dose i is ineffective

$\vdots$

$H_{02}^{\downarrow}$: Doses $3, \ldots, k$ are effective but dose 2 is ineffective

Statistically, the null hypotheses are:

$$\begin{aligned}
&H_{0k}^{\downarrow}: \mu_k \le \mu_1 + \Delta\\
&H_{0(k-1)}^{\downarrow}: \mu_{k-1} \le \mu_1 + \Delta < \mu_k\\
&\vdots\\
&H_{0i}^{\downarrow}: \mu_i \le \mu_1 + \Delta < \min\{\mu_{i+1}, \ldots, \mu_k\}\\
&\vdots\\
&H_{02}^{\downarrow}: \mu_2 \le \mu_1 + \Delta < \min\{\mu_3, \ldots, \mu_k\}\,.
\end{aligned} \tag{5}$$

For any integer i, if $H_{0j}^{\downarrow}$, $j = i, \ldots, k$, are all rejected, then the logical inference is doses $i, \ldots, k$ are all efficacious: $\mu_j > \mu_1 + \Delta$, $j = i, \ldots, k$.

Let MED^U denote the smallest integer M such that $H_{0j}^{\downarrow}$, $j = M, \ldots, k$, are all rejected (with the understanding that if no $H_{0i}^{\downarrow}$ is rejected, then $\mathrm{MED}^U = k+1$). Suppose, for each individual null hypothesis $H_{0i}^{\downarrow}$, the probability of rejecting it if it is true is not more than α, then MED^U will be a $100(1-\alpha)\%$ upper confidence bound on MED. This can be seen quite explicitly as follows. Let M^* be the smallest integer between 2 and k so that $\mu_i - \mu_1 > \Delta$ for all i, $i \ge M^*$, i.e., dose M^* is the true MED. If no such integer exists, then (for convenience) let M^* be $k+1$. Either all null hypotheses are false (if $M^* = 2$), or exactly one null hypothesis, $H_{0(M^*-1)}^{\downarrow}$, is true (if $M^* > 2$). If all null hypotheses are false, then it is correct to infer any set of doses to be efficacious; no error is made regardless of what MED^U is. If $M^* > 2$, then an error is made if $M < M^*$, inferring doses $M, \ldots, k$ to be efficacious but $M < M^*$. This can only happen if $H_{0(M^*-1)}^{\downarrow}$ is rejected. Since the probability of rejecting a true $H_{0(M^*-1)}^{\downarrow}$ is no more than α, the probability of an incorrect MED^U inference is no more than α. The interesting thing is, in simultaneously testing $H_{0j}^{\downarrow}$, $j = 1, \ldots, k$, no multiplicity adjustment is needed to control the probability of rejecting any true null hypothesis. This is because at most one null hypothesis can be true. We thus test each $H_{0j}^{\downarrow}$, $j = 2, \ldots, k$, at level α.

Level α tests for each $H_{0i}^{\downarrow}$, $i = 2, \ldots, k$, are of course not unique. Note, however, a level-α test for

$$H_{0i}: \mu_i \le \mu_1 + \Delta \tag{6}$$

is also a level-α test for

$$H_{0i}^{\downarrow}: \mu_i \leq \mu_1 + \Delta < \min\{\mu_{i+1}, \ldots, \mu_k\}\,.$$

So the simplest level-α test for $H_{0i}^{\downarrow}$ is to use a one-sided two-sample size α t-test comparing μ_i with μ_1 for each $H_{0i}^{\downarrow}$. With this choice of test for $H_{0i}^{\downarrow}$, $i = 2, \ldots, k$, and executing them in a stepwise fashion for MED inference, Hsu and Berger (1999) showed in fact one can attach confidence bounds to the inference:

$\boxed{\text{Step 1}}$

$$\begin{array}{ll} \text{If} & \hat{\mu}_k - \hat{\mu}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_1} > \Delta\,, \\ \text{then} & \text{infer } \mu_k - \mu_1 > \Delta \text{ and go to Step 2}\,; \\ \text{else} & \text{infer } \mu_k - \mu_1 > \hat{\mu}_k - \hat{\mu}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_k + 1/n_1} \text{ and stop}\,. \end{array}$$

$\boxed{\text{Step 2}}$

$$\begin{array}{ll} \text{If} & \hat{\mu}_{k-1} - \hat{\mu}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_{k-1} + 1/n_1} > \Delta\,, \\ \text{then} & \text{infer } \mu_{k-1} - \mu_1 > \Delta \text{ and go to Step 3}\,; \\ \text{else} & \text{infer } \mu_{k-1} - \mu_1 > \hat{\mu}_{k-1} - \hat{\mu}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_{k-1} + 1/n_1} \text{ and stop}\,. \end{array}$$

$\vdots$

$\boxed{\text{Step } k-1}$

$$\begin{array}{ll} \text{If} & \hat{\mu}_2 - \hat{\mu}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_2 + 1/n_1} > \Delta\,, \\ \text{then} & \text{infer } \mu_2 - \mu_1 > \Delta \text{ and go to Step } k\,; \\ \text{else} & \text{infer } \mu_2 - \mu_1 > \hat{\mu}_2 - \hat{\mu}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_2 + 1/n_1} \text{ and stop}\,. \end{array}$$

$\boxed{\text{Step } k}$

$$\text{Infer } \min_{i=2,\ldots,k} \mu_i - \mu_1 > \min_{i=2,\ldots,k}\{\hat{\mu}_i - \hat{\mu}_1 - t_{\alpha,\nu}\hat{\sigma}\sqrt{1/n_i + 1/n_1}\} \text{ and stop.}$$

Note the Step k inference results from pivoting Intersection-Union Tests (IUT) for

$$H_{0\Delta}^{\downarrow}: \min_{i=2,\ldots,k} \mu_i = \mu_1 + \Delta, \quad \Delta \in (0, \infty) \tag{10}$$

A common misconception is the validity of this stepdown method depends on an assumption that the true response is monotonically non-decreasing as dose increases. Actually, the stepwise method is valid without any assumption on the response curve. This is because the union of the null hypotheses

$$H_{0i}^{\downarrow}: \mu_i \leq \mu_1 + \Delta < \min\{\mu_{i+1}, \ldots, \mu_k\}$$

$i = 2, \ldots, k$, together with

$$H_{0\Delta}^{\downarrow}: \min_{i=2,\ldots,k} \mu_i = \mu_1 + \Delta, \quad \Delta \in (0, \infty)$$

exhaust the entire parameter space of all possible response curves.

To apply this stepwise method using SAS, assuming the doses are labeled $1, \ldots, k$ with dose 1 being the control against which other doses should be compared, one can execute the code

```
PROC GLM;
CLASS dose;
MODEL response = dose;
LSMEANS dose/PDIFF=CONTROLU('1') ADJUST=T CL ALPHA=.05;
```

and then compares the individual (not simultaneous) lower confidence bounds to Δ in a stepwise fashion, as illustrated below.

3.2.4 Dose finding study example (continued)

For the dose finding study example from Section 2, again suppose a dose is considered effective if the mean response is higher than the zero-dose by $\Delta = 40$. Then individual lower confidence bounds (without multiplicity adjustment) from SAS are as follows.

```
 i    j      Difference          95% Confidence Limits for
             Between Means       LSMean(i) - LSMean(j)

 2    1     -1.666707          -9.165839        Infinity
 3    1      2.151515          -5.347617        Infinity
 4    1      7.856362           0.357230        Infinity
 5    1     14.974230           7.475098        Infinity
 6    1     32.393775          24.894643        Infinity
 7    1     48.884348          41.385216        Infinity
 8    1     47.834935          40.335803        Infinity
 9    1     47.927485          40.428353        Infinity
10    1     50.667818          43.168686        Infinity

NOTE: To ensure overall protection level, only probabilities
      associated with pre-planned comparisons should be used.
```

Based on this output, the following stepwise procedure is used for MED inference:

$\boxed{\text{Step 1}}$

Is 43.168686 > 40? Yes, infer $\mu_{10} - \mu_1 > 40$ and go to Step 2 .

$\boxed{\text{Step 2}}$

Is 40.428353 > 40? Yes, infer $\mu_9 - \mu_1 < 0$ and go to Step 3 .

$\boxed{\text{Step 3}}$

Is 40.335803 > 40? Yes, infer $\mu_8 - \mu_1 < 0$ and go to Step 4 .

$\boxed{\text{Step 4}}$

Is 41.385216 > 40? Yes, infer $\mu_7 - \mu_1 < 0$ and go to Step 5 .

$\boxed{\text{Step 5}}$

Is 32.393775 > 40? No, infer $\mu_6 - \mu_1 > 32.393775$ and stop .

Since dose group 7, 8, 9 and 10 are inferred to be effective, the MED is estimated to be the dosage corresponding to dose group 7, which is 3.0 mg/kg in this example.

3.2.5 Choice of dose response analysis method

If the dose levels giving the higher sample responses are the ones tested early in the steps of the stepwise method, then the stepwise method will infer more doses as efficacious than Dunnett's method. On the other hand, if the dose levels giving the higher sample responses are not the ones tested early in the steps of the stepwise method, then Dunnett's method will infer more doses as efficacious than the stepwise method.

There is nothing sacred about testing dose k then dose $k-1$ then dose $k-2\dots$ in that sequence; any *pre-specified* sequence is valid. One could specify the sequence: test dose 3 then dose 4 then dose 5 then dose 2, for example, if an inverted-U-shaped response curve is anticipate. One should, of course, choose a sequence of doses so that the resulting set of doses inferred to be efficacious will be contiguous.

Alternative analyses methods exist which either restrict the parameter space or make further assumptions than considered in this section. Trend tests, for example, are often applied to assess the null hypothesis $H_0: \mu_1 = \ldots = \mu_k$ of no dose response effect against the restricted alternative hypothesis $H_1: \mu_1 \leq \ldots \leq \mu_k$, $\mu_1 < \mu_k$. While such tailored methods can be very powerful to detect a dose response relationship different from a flat model if H_1 in fact holds, problems arise if the true parameter vector lies outside $H_0 \cup H_1$ and the resulting p-values and confidence intervals become difficult to interpret. We refer to Robertson et al. (1988), Hothorn (2006), Bretz (2006) and the references therein for details.

4 Modeling Approaches

The modeling approach to dose finding is based on an assumed functional relationship between the clinical endpoint and the dose, treated as a continuous variable, according to a pre-specified parametric model, such as a logistic or an $E_{\max}$ model (see Table 2). The model fitted to the observed data is used to test for the presence of dose response and to estimate an adequate dose(s) to achieve a desired response using, for example, inverse regression techniques. One of the main appeals of this approach is that it provides flexibility in investigating the effect of doses not used in the actual study. However, the validity of its conclusions will highly depend on the correct choice of the dose response model, which is of course *a priori* unknown. This creates a dilemma in practice, because, within the regulated environment in which drug development takes place, it is required to have the analysis methods (including the choice of the dose response model) defined at the study design stage. We will revisit this issue in the next section, when we discuss the MCP-Mod approach.

The general framework adopted in this section is that a response Y (which can be an efficacy or a safety variable) is observed for a given set of *parallel* groups of patients corresponding to doses

Table 2 A selection of frequently used dose response models.

Model	$f(d,\boldsymbol{\theta})$	$f^0(d,\boldsymbol{\theta}^0)$
Linear	$E_0 + \delta d$	d
Linear log-dose	$E_0 + \delta \log(d+c)$	$\log(d+c)$
Exponential	$E_0 + E_1[\exp(d/\delta) - 1]$	$\exp(d/\delta) - 1$
$E_{\max}$	$E_0 + E_{\max} d/(ED_{50} + d)$	$d/(ED_{50} + d)$
Sigmoid $E_{\max}$	$E_0 + E_{\max} d^h/(ED_{50}^h + d^h)$	$d^h/(ED_{50}^h + d^h)$
Logistic	$E_0 + E_{\max}/\{1 + \exp[(ED_{50} - d)/\delta]\}$	$1/\{1 + \exp[(ED_{50} - d)/\delta]\}$
Quadratic	$E_0 + \beta_1 d + \beta_2 d^2$	$d + (\beta_2/\lvert\beta_1\rvert)d^2$
Beta	$E_0 + E_{\max} B(\alpha,\beta)(d/D)^\alpha (1-d/D)^\beta$	$(d/D)^\alpha (1-d/D)^\beta$

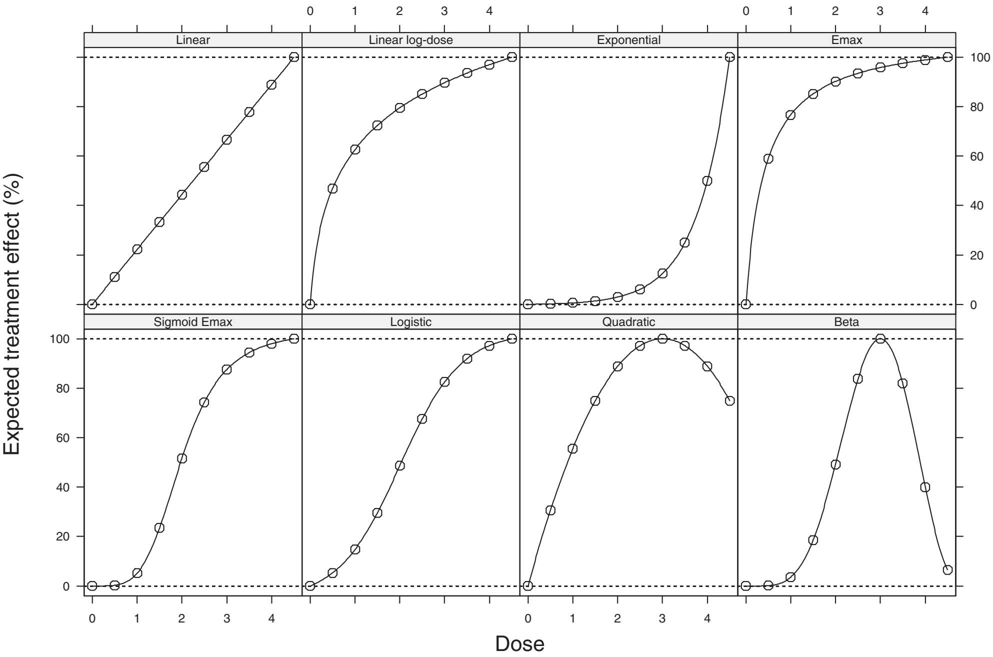

Figure 3 Dose response profiles corresponding to the models in Table 2. Open dots indicate the responses at the dose levels used in the Ruberg example of Section 2. Response is represented as the percentage of the maximum treatment effect achievable within the dose range.

$d_2, d_3, \ldots, d_k$ plus placebo d_1, for a total of k arms. The model is then specified as

$$Y_{ij} = f(d, \boldsymbol{\theta}) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i\,, \tag{7}$$

where $f(.)$ is parameterized by a vector of parameters $\boldsymbol{\theta}$ and ε_{ij} is the error term for patient j within dose group i. Most dose response models used in practice can be expressed as a location-scale family of a *standardized model*, f^0, such that $f(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(d, \boldsymbol{\theta}^0)$. This decomposition is often useful in practice; for example, in fitting nonlinear regression models, starting estimates are only required for the standardized model parameters $\boldsymbol{\theta}^0$.

Table 2 lists a (non-exhaustive) selection of models frequently used to represent dose response relationships, together with their respective standardized versions. Note that some of the models are linear in their respective parameters (e.g., linear and quadratic), while others depend non-linearly on at least some of its parameters (e.g., $E_{\max}$ and logistic). The quadratic model described in the table is assumed to be "umbrella-shaped" with $\beta_2 < 0$. For the beta model, $B(\alpha, \beta) = (\alpha + \beta)^{\alpha+\beta}/(\alpha^\alpha \beta^\beta)$. A graphical display of each these models is shown in Figure 3, using the dose design of Ruberg's example described in Section 2. Pinheiro et al. (2006a) described several linear and nonlinear regression dose response models commonly used in practice, including clinical interpretations for the associated model parameters. In broader context, Ratkowsky (1989) summarized many further non-linear regression models, some of which are well applicable to clinical practice.

4.1 Fitting dose response models

The modeling approach to dose finding requires the estimation of the model parameters $\boldsymbol{\theta}$ in the assumed dose response model. Under the assumption of independent, identically distributed errors ε_{ij}

adopted here, ordinary least squares (OLS) estimates that minimize the residual sum of squares $\sum_{i=1}^{k}\sum_{j=1}^{n_i}|Y_{ij}-f(\boldsymbol{\theta},d_i)|^2$ are typically used. In the case of nonlinear dose response models, *nonlinear least squares* algorithms are needed to estimate $\boldsymbol{\theta}$. The most popular of these is the Gauss–Newton algorithm (Bates and Watts, 1988; Seber and Wild, 1989), which is an iterative procedure consisting of solving, until convergence, a sequence of linear least squares problems based on a local approximation of the nonlinear model. Such iterative algorithms typically require a starting point, the so-called *initial values*, for the parameters. Methods for deriving initial estimates for nonlinear models are discussed in Bates and Watts (1988).

The Gauss–Newton algorithm for nonlinear least squares is implemented in mainstream statistical software packages. The functions `nls` (Bates and Chambers, 1992) and `gnls` (Pinheiro and Bates, 2000) implement it in S-PLUS and R, while `PROC NLIN` (Freund and Littell, 2000) implements it in SAS. Examples on the use of these functions and procedure can be found in the references given above.

As an illustration, we use the data from Ruberg (1995b), presented in Section 2. As mentioned there, a logistic model seems to describe well the observed dose response shape. Closer inspection of Figure 2 provides initial values for the parameters in the logistic model. For example, the basal level E_0 is around 25 and the maximum increase over it, $E_{\max}$ is around 50. Furthermore, the inflection point ED_{50} is around 2.5 and about 75% of the maximum effect is attained at about dose 3, which suggests an initial value of 0.5 for the scale parameter δ. Assuming the data is available in S (S-PLUS or R) as a `data.frame` object `ruberg` with columns `resp` and `dose`, the following code can be used to fit the logistic model.

```
> fmRub <- nls(resp ~ e0 + eM/(1 + exp((ed50 - dose)/delta)), ruberg,
                  start = list(e0 = 25, eM = 50, ed50 = 2, delta = 0.5))
```

The parameter estimates, provided in the output below, are in close agreement with the initial values.

```
> summary(fmRub)
. . .
Parameters:
            Value      Std. Error     t value
   e0   25.754800    1.9612700    13.13170
   eM   49.907100    3.0056100    16.60470
 ed50    2.257130    0.0809226    27.89240
delta    0.337437    0.0718254     4.69802
Residual standard error: 7.58025 on 56 degrees of freedom
. . .
```

Note that the high t-value for the $E_{\max}$ parameter (represented by `eM` in the output) indicates that there is statistically significant evidence of dose response in the data. Alternatively, we can compare the fit above with one assuming a flat dose response, via a Wald F-test.

```
> fmFlat <- lm(resp ~ 1, rub)
> anova(fmRub, fmFlat)
. . .
   F Value Pr(F)
1
2  118.721  0
```

As expected, very significant evidence of PoC. Figure 4 shows the fitted logistic curve with 95% confidence bands (pointwise).

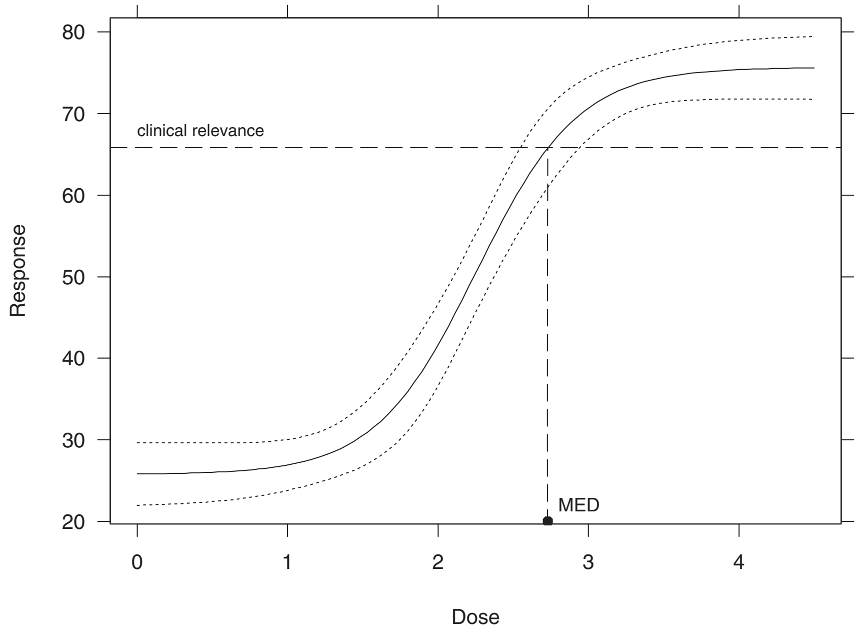

Figure 4 Fitted logistic regression model corresponding to Ruberg's example of Section 2, with 95% pointwise confidence bands. Clinically relevant improvement of $\Delta = 40$ over placebo and the estimated smallest dose producing it (i.e., the MED) are indicated.

4.2 Estimation of target doses

Once the dose response model has been successfully fitted to the data and PoC established, one may proceed to estimate the target dose(s) of interest. We restrict ourselves here to two types of target dose: the MED introduced previously and the ED$_p$, defined as the smallest dose that gives 100p% of the maximum effect achievable in the observed dose range. For example, the ED$_{95}$ is the smallest dose producing 95% of the maximum effect. For a given clinically relevant effect Δ, the MED associated with a model $f(d, \boldsymbol{\theta})$ is defined as

$$\text{MED} = \text{argmin}_{d \in (d_1, d_k]} \{f(d, \boldsymbol{\theta}) \geq f(d_1, \boldsymbol{\theta}) + \Delta\} . \qquad (8)$$

Note that the MED need not exist, as no dose in $(d_1, d_k]$ may produce an improvement of Δ over placebo. We restrict the MED to lie within the interval $(d_1, d_k]$ in order to avoid problems arising from extrapolating beyond the dose range under investigation. Note that definition (8) is different from the definition of the MED used in Section 3.2. Here we assume the MED estimate to take any value within the continuous dose range $(d_1, d_k]$ and is thus not confined to the dose levels under investigation. In contrast to Section 3.2 we also drop the restriction that all doses larger than the MED have to be effective as well.

Letting $g(d, \boldsymbol{\theta}) = f(d, \boldsymbol{\theta}) - f(d_1, \boldsymbol{\theta})$, the ED$_p$ can be is similarly defined as

$$\text{ED}_p = \text{argmin}_{d \in (d_1, d_k]} \{g(d, \boldsymbol{\theta}) / g(d_{\max}, \boldsymbol{\theta}) \geq p\}, \quad d_{\max} = \text{argmax}_{d \in (d_1, d_k]} g(d, \boldsymbol{\theta})$$

Unlike the MED, the ED$_p$ always exists.

We consider initially an estimate for the MED, proposed by Bretz et al. (2005). Denote by L_d the lower $1 - \gamma$ confidence bound for the predicted value P_d at dose d based on the fitted model $f(d, \widehat{\boldsymbol{\theta}})$, as computed, for example, by the `nls` function in S. As discussed in Bretz et al. (1005), γ does not need to be specified with strong control of type I error rates in mind, but should, nevertheless, be set reasonably small in order to avoid interpretation problems with the final estimate of the MED. The following estimate for the MED is then proposed.

$$\widehat{\text{MED}} = \text{argmin}_{d \in (d_1, d_k]} \{P_d \geq P_{d_1} + \Delta, L_d > P_{d_1}\} .$$

An estimate for the ED$_p$ can be similarly determined. Let l_d denote the lower $1 - \gamma$ confidence bound for the predicted value p_d of $g(d, \boldsymbol{\theta})$. The estimate for ED$_p$ is then defined as

$$\widehat{\text{ED}_p} = \text{argmin}_{d \in (d_1, d_k]} \{p_d / p_{d_{\max}} \geq p, l_d > 0\}, \quad d_{\max} = \text{argmax}_{d \in (d_1, d_k]} \, p_d . \tag{9}$$

As an illustration, we consider again Ruberg's example. Assume that the clinically relevant difference for that particular trial was $\Delta = 40$. Because the predicted response at $d_1 = 0$ is 25.8168, the estimated MED, using $\gamma = 0.025$, is given by the smallest dose with a predicted value larger than or equal to 65.8168 and a lower limit for the corresponding 95% confidence interval greater than 25.8168. Because of the relatively narrow confidence bounds for the fitted model (see Figure 4), the $\widehat{\text{MED}}$ is determined by the first condition, being equal to 2.731, which fits nicely with the estimate of 3.0 mg/kg from Section 3.2.4. The derivation of the $\widehat{\text{MED}}$ for this example is illustrated graphically in Figure 4. The estimated ED$_{90}$ for the Ruberg data, based on (9), is 2.995.

Once an estimate for the MED, or the ED$_p$, is obtained using the methods above, it is important to assess its precision, for which a confidence interval is generally the most useful tool. Bootstrap methods can be used to derive such a confidence interval. Nonparametric bootstrap methods, for which responses are sampled with replacement within each dose level (preserving the sample sizes per dose), are easy to implement in S and provide reliable confidence intervals. Applying this approach to Ruberg's example with 1000 non-parametric bootstrap samples and the same assumptions as before for estimating the MED, produces a 95% confidence intervals for the MED of $[2.518, 2.959]$, suggesting reasonably good precision. The 95% bootstrap confidence interval for the ED$_{90}$ is $[2.691, 3.433]$.

4.3 Extensions and alternative modeling approaches

The focus of the methods discussed in this section has been on parallel group designs, with normally distributed data, for which a parametric model was assumed, and the only covariate used in the model was dose. Different types of designs (e.g., crossovers) and responses (e.g., binary) are observed in clinical practice and require extensions, or alternatives, for the methods discussed. Some of those are briefly touched upon below.

Baseline covariates

The inclusion of baseline covariates in dose response models is easily done when only the base term E_0 is allowed to depend on the covariates. The modeling becomes more complex when other terms (e.g., $E_{\max}$) are also allowed to depend on the covariates. The latter would be associated with treatment × covariate interactions, which are typically only included as secondary analysis in clinical trials (e.g., ANCOVA models). When the inclusion of covariates in the model are restricted to E_0, the methods described in this section are easily generalized.

Non-Gaussian responses

Even though the basic ideas and modeling principles discussed here apply to most types of response variables, different models and estimation methods will be required to properly analyze non-Gaussian data. Generalized linear models (McCullagh and Nelder, 1989) and generalized nonlinear models (Turner and Firth, 2007) provide useful tools for handling most types of non-Gaussian data encountered in clinical trials. Both of these approaches have been implemented in R (`glm` and `gnm` functions, respectively).

Correlated data
When repeated measures data are used for modeling, such as, for example, in crossover trials, the assumption of independence among observations typically will not hold. Mixed-effects models (Pinheiro and Bates, 2000) offer a flexible alternative for model-based dose finding in these cases. Linear and non-linear mixed-effects models can be used with Gaussian data, while generalized linear mixed-effects models can handle non-Gaussian data. The basic principles of model-based dose-finding described in the previous sections also apply to mixed-effects models, but the methods and estimates are more complex. Reliable software for fitting such models is available in mainstream packages, such as SAS, S-PLUS, and R.

Non-parametric models
An alternative to having to assume a parametric model prior to the start of the trial is to use non-parametric modeling to estimate the dose response after the data becomes available. Although more flexible and robust to model misspecification, this approach can be inefficient when there is good prior knowledge about the approximate shape of dose response profile. Bornkamp et al. (2007) considered a non-parametric dose response modeling approach based on local polynomial fits (Loader, 1999), reporting simulation results comparing their method to other model-based dose-finding approaches.

Dose-exposure-response models
Pharmacometricians tend to approach dose response modeling by decomposing it into two submodels (Holford, 2006). The first is based on pharmacokinetic (PK) modeling of the concentration curve, being aimed at establishing the relationship between dose and some measure of exposure, such as the area under the concentration curve (AUC), or the maximum concentration (C_{max}). The second model relates the exposure to the response, via a pharmacodynamic (PD) model (Sun and Fadiran, 2007). The appeal of this approach is that inter-patient variation in dose response is often associated with differences in exposure among patients for the same dose. Therefore, it is expected that the exposure-response model will be estimated with higher precision than the dose response model. However, if the dose-exposure response is highly variable, the benefit of this approach for selection of a single dose for Phase III still needs to be further evaluated.

Bayesian methods
Bayesian methods can be used with the model-based methods described in this section, as well as with any of the extensions mentioned above – and, in fact, they have been used quite extensively in the context of dose finding. Estimates and decision rules (e.g., PoC test) need to be adapted to the Bayesian context, in particular, to rely on the posterior distribution of parameters given the observed data. Bayesian dose-finding methods offer additional flexibility and, in some cases, ease of interpretation compared to frequentist approaches, but come with the usual caveats of Bayesian inference, including the need to specify priors for all model parameters and increased computational complexity. For examples and discussion of Bayesian dose-finding methods see Berry et al. (2001) and Bornkamp et al. (2007).

5 Combining Multiple Comparisons and Modeling Techniques

As seen from the previous sections, multiple comparison procedures provide solid inferences, which can be used, for example, to test for a significant dose response signal (i.e., PoC) or whether the effect at certain dose levels differ significantly from the placebo response. However, when using standard multiple comparison procedures, any inference is restricted to the distinct dose levels being administered in a given trial. Modeling approaches provide more flexibility by considering the dose as a continuous variable. Any dose within the range under investigation can potentially be declared as target dose estimate. The drawback of modeling approaches is their dependence on the regression model, resulting in model uncertainty, which is often underestimated. This is because the dose response profile is typically unknown prior to a clinical study. Figure 3 displays a non-exhaustive set of dose response profiles, which are particularly relevant in the context of clinical dose finding studies. Fitting a working model f, say, provides a parameter estimate $\hat{\boldsymbol{\theta}}$ conditional on f. Consequently, the variance of the estimate $\hat{\boldsymbol{\theta}}$ is $\text{var}(\hat{\boldsymbol{\theta}}\,|f)$, which can be substantially smaller than the unconditional variance $\text{var}(\hat{\boldsymbol{\theta}})$ typically meant to be reported in practice. Ignoring model uncertainty thus can lead to highly undesirable effects, see Chatfield (1995), Draper (1995), and Hjorth (1994).

Hybrid dose finding methods that combine principles of multiple comparisons with modeling techniques have recently been investigated to overcome some of the shortenings of applying either approach alone. An early reference on these methods is Tukey et al. (1984), who recognized that the power of standard hypotheses tests to detect a dose response signal depends critically on the (unknown) dose response relationship. They proposed to simultaneously use several trend tests and subsequently to adjust the resulting p-values for multiplicity. Bretz et al. (2005) proposed an extension of this methodology, denoted MCP-Mod, which provides the flexibility of modeling for dose estimation, while preserving the robustness to model misspecification associated with multiple comparison procedures. Practical considerations regarding the implementation of this methodology were discussed by Pinheiro et al. (2006b). Extensions to Bayesian methods estimating or selecting the dose response curve from a sparse dose design have also been investigated (Neal, 2006; Wakana et al., 2007). Optimal designs with respect to the number of different dose levels k, the location of the dose levels $d_2, \ldots, d_{k-1}$ and the proportions of patients allocated to the individual dose levels have been derived by Dette et al. (2008). In the following we review the MCP-Mod approach proposed by Bretz et al. (2005) and illustrate the method with the example from Ruberg (1995b). Software implementations of MCP-Mod are available as add-on packages both in S-PLUS (http://csan.insightful.com) and R (www.r-project.org, Bornkamp et al. (2008)).

5.1 The MCP-Mod approach

The central idea of the MCP-Mod approach is to use a set $\mathcal{M}$ of candidate dose response models to cover the possible shapes anticipated for the dose response relationship. For each model shape a contrast is derived that is optimal for detecting the relevant shape. Multiple comparison procedures are applied to the resulting contrast tests. If any of the contrast tests is significant, a best model (or a combination of models) is chosen out of the set of significant models and used for estimating the target dose employing modeling techniques.

To formalize the ideas, assume that a set $\mathcal{M}$ of M parameterized candidate models is given, with corresponding model functions $f_m(d, \boldsymbol{\theta}_m)$, $m = 1, \ldots, M$ and prior guesses for the parameters of the standardized models $\boldsymbol{\theta}_m^0$, determining the model shapes (recall equation (7) for the definition of the full and the standardized models). Each of the dose response shapes in the candidate set is now tested using a single contrast test, with contrast coefficients chosen to maximize the power of the contrast tests introduced further below. The single contrast test for detecting the m-th model shape is defined by

$$T_m = \frac{\sum_{i=1}^{k} c_{mi} \bar{Y}_i}{S\sqrt{\sum_{i=1}^{k} c_{mi}^2/n_i}}, \qquad m = 1, \ldots, M\,, \tag{10}$$

where $S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(N-k)$ and $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{mk})'$ is the optimal contrast vector for detecting model shape m. Letting $\boldsymbol{\mu}_m^0 = (\mu_{m1}^0, \ldots, \mu_{mk}^0)' = (f_m^0(d_1, \boldsymbol{\theta}_m^0), \ldots, f_m^0(d_k, \boldsymbol{\theta}_m^0))'$, the associated null hypothesis to be tested is $H_{0m}\colon \boldsymbol{c}_m \boldsymbol{\mu}_m^0$. Every single contrast test thus translates into a decision procedure whether a given dose response shape is statistically significant, based on the observed data. It can be shown that the optimal contrast coefficients depend only on the standardized model parameters $\boldsymbol{\theta}_m^0$. The ith entry of the optimal contrast $\boldsymbol{c}_m$ for detecting shape m is proportional to

$$n_i(\mu_{mi}^0 - \bar{\mu}), \quad i = 1, \ldots, k, \tag{11}$$

Table 3 Candidate dose response models for the example from Ruberg (1995b). All models are normalized such that the placebo effect is 25 and the maximum effect over placebo is 50.

Model	Specification
linear	25 + (50/4.5) d
$E_{\max}$	$25 + 72.2222d/(2+d)$
logistic	$24.9992 + 50.0085/(1+\exp(-(d-2.5)/0.2276))$
exponential	$25 + 3.4432(\exp(d/1.6410) - 1)$

where $\bar{\mu} = N^{-1}\sum_{i=1}^{k} \mu_{mi}^{0} n_i$ (Bretz et al., 2005; Bornkamp, 2006). A unique representation of the optimal contrast can be obtained by imposing the regularity condition $\sum_{i=1}^{k} c_{mi}^2 = 1$.

The final test statistic $T_{\max} = \max_m T_m$ is the maximum of all single contrast tests. Its distribution can be derived from the joint distribution of $T = (T_1, \ldots, T_M)'$. Under the null hypothesis of no dose response effect (i.e., $\mu_{d_1} = \ldots = \mu_{d_k}$) and the distributional assumptions stated in (7), $T_1, \ldots, T_M$ jointly follow a central multivariate t distribution with $N-k$ degrees of freedom and correlation matrix $R = (\varrho_{ij})$, where

$$\varrho_{ij} = \frac{\sum_{l=1}^{k} c_{il}c_{jl}/n_l}{\sqrt{\sum_{l=1}^{k} c_{il}^2/n_l \sum_{l=1}^{k} c_{jl}^2/n_l}}\,. \tag{12}$$

Numerical integration routines, such as those implemented in the R package *mvtnorm* (Hothorn et al., 2001; Genz and Bretz, 2002) can be used to compute multiplicity adjusted p-values and critical values. PoC can hence be established if $T_{\max} \geq q_{1-\alpha}$, where $q_{1-\alpha}$ is the multiplicity adjusted critical value from the multivariate t distribution. Furthermore, all dose response shapes with contrast test statistics larger than $q_{1-\alpha}$ can be declared statistically significant at level α. These models then form a reference set $\mathcal{M}^*$ of significant dose response models. If no candidate model is statistically significant, the reference set $\mathcal{M}^*$ is empty and the procedure stops, indicating that a dose response relationship can not be established from the observed data (i.e., no PoC).

If at least one model shape is significant, a suitable dose response model is selected from the reference set $\mathcal{M}^*$. The selection of the model can be based on the minimum p-value or any other model selection criteria, like AIC or BIC. The selected dose response model is used to fit the data by maximizing the likelihood of the model with respect to its parameters $\boldsymbol{\theta}$. For non-linear models iterative optimization techniques have to be used, see Section 4.1. Once the dose response model has been successfully fitted to the data and PoC established, the final step of the MCP-Mod approach is to estimate the target dose(s) of interest based on the fitted model, using the methods described in Section 4.2.

5.2 Revisiting the dose finding example

To illustrate the MCP-Mod methodology reviewed in Section 5.1, we revisit Ruberg's example from Section 2 and perform a post-hoc analysis to test for the presence of a significant dose response signal and, if so, to estimate the smallest dose achieving a clinically relevant effect. We begin by considering four potential dose response models to form our candidate set, described in Table 3 with respective parameter values and displayed in Figure 5.

Based on the model specifications from Table 3 we can calculate the optimal contrasts using (11). For example, for the $E_{\max}$ model we have the standardized model $f^0(d, \boldsymbol{\theta}^0) = d/(2+d)$ with $\boldsymbol{\mu}^0 = (0, 0.2, 0.333, 0.429, 0.5, 0.556, 0.6, 0.636, 0.667, 0.692)'$ and $\bar{\mu} = 0.461$, so that after normaliza-

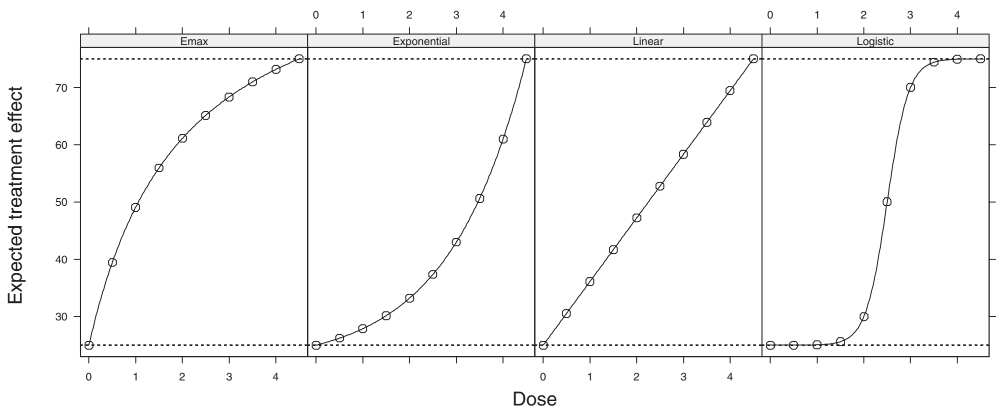

Figure 5 Dose response profiles corresponding to the candidate models in Table 3.

tion $c = (-0.684, -0.388, -0.19, -0.049, 0.057, 0.14, 0.206, 0.26, 0.305, 0.343)'$. We similarly calculate the optimal contrasts for the remaining models and obtain

```
        linear      emax     logistic    exponential
0       -0.495    -0.684      -0.317        -0.319
0.5     -0.385    -0.388      -0.317        -0.295
1       -0.275    -0.190      -0.316        -0.262
1.5     -0.165    -0.049      -0.308        -0.216
2       -0.055     0.057      -0.246        -0.155
2.5      0.055     0.140       0.035        -0.072
3        0.165     0.206       0.317         0.041
3.5      0.275     0.260       0.379         0.194
4        0.385     0.305       0.386         0.401
4.5      0.495     0.343       0.387         0.683
```

Note that by construction $\sum_i c_i = 0$ for each of the previous four contrasts. The plot of the contrast coefficients (not shown here) indicates that they are indeed mimicking closely the candidate models from Figure 5. The associated correlation matrix (with rows and columns in the same order as the contrasts above)

$$R = \begin{pmatrix} 1 & 0.946 & 0.929 & 0.937 \\ 0.946 & 1 & 0.819 & 0.789 \\ 0.929 & 0.819 & 1 & 0.885 \\ 0.937 & 0.789 & 0.885 & 1 \end{pmatrix}$$

Table 4 Contrast test statistics and information criteria for the candidate dose response models in the example from Ruberg (1995b).

Statistic	Model			
	Linear	$E_{\max}$	Logistic	Exponential
t-test	20.392	18.869	21.048	18.550
AIC	447.65	449.60	419.20	NA
BIC	453.93	457.97	429.67	NA

calculated from (12) indicates that the contrasts are highly correlated, the minimum correlation is about 0.8. Consequently, the associated critical value $q_{0.975} = 2.24$ from the multivariate t distribution is substantially smaller than the value of 2.6 resulting from a simple Bonferroni adjustment.

The contrast test statistics, included in Table 4, were computed by applying (10) to the observed mean values and standard deviations from Table 1. All four t-test statistics are highly significant, giving a clear indication of the presence of a dose response signal. The maximum contrast test statistic $T_{\max} = 21.048$ is associated with the logistic model, suggesting the greater adequacy of this model. This is further reinforced by the AIC and BIC values (obtained from the corresponding fitted models), which both clearly indicate the superiority of the logistic model fit. The information criteria provide considerably stronger evidence in favor of the logistic model, compared to the contrast test statistics. Convergence was not attained for the exponential model, hence the NA values for its AIC and BIC in Table 4.

The logistic model was therefore the one selected for the final dose estimation step of MCP-Mod, leading to the same MED estimate and conclusions presented in Section 4.1. Note that in contrast to Section 4.1 the MCP-Mod methodology does account for model uncertainty by considering an initial set of candidate models in a rigid testing environment. To further illustrate this point, we calculated a confidence interval for the MED using non-paramateric bootstrap resampling, as described in Section 4.1, but applying the full MCP-Mod algorithm to each resample. This allows different dose response models from the candidate set to be chosen for the MED estimation in each resample, thus incorporating model uncertainty into the calculations. The resulting 95% confidence interval, $[2.51, 2.97]$, was very similar to the one presented in Section 4.1 (based on the logistic model alone). This is explained by the very strong signal for the logistic model in the data, which led to this model being selected in 99.2% of the 1000 bootstrap samples (the linear model was used in the remaining 0.8% of the cases), making it in essence equivalent to the bootstrap using the logistic model only, as done in Section 4.1. In applications where the signal for one particular dose response shape is not so prevalent in the data as in Ruberg's example (likely to happen when fewer doses are used, for example), model uncertainty will play a more important role and will lead to wider (and more adequate) confidence intervals for the MED based on the MCP-Mod bootstrap, compared to the single-model approach.

6 Adaptive Dose Finding Studies

Adaptive dose finding designs have attracted considerable interest recently, since they offer one possibility to improve on the previously described methods. These designs offer efficient ways to learn about the dose response through repeated looks at the data being accrued during the conduct of a clinical trial. This interim information can be used to guide decision making and for example optimize the allocation of the patients based on the collected information. In particular, doses can be discontinued due to futility or safety reasons and even an early stop of the whole study is possible. Adaptive dose finding methods can be applied within a regular dose finding study. But they can also support innovative trial designs, such as combining proof-of-concept studies with dose finding into a single trial. The continuation of a dose finding trial into a confirmatory stage through a seamless design is a further opportunity to increase information on the correct dose earlier in development, and thus reduce the total duration of the clinical development program. Accordingly, we first review confirmatory adaptive designs which strongly control the overall type I error rate. An overview of adaptive dose finding methods aimed at Phase II trials is given afterwards.

Confirmatory adaptive designs extend the classical group sequential procedures (Jennison and Turnbull, 2000) and offer a high level of flexibility during the conduct of a clinical trial, including the possibility to select dose levels at an interim analysis. Several adaptive designs methods have been proposed. They differ with respect to how the evidence from different stages of the trial is combined. In all cases it is decided based on the unblinded data collected up to the interim analysis, whether the trial is continued (conducting the next stage) or not (early stopping: either due to futility or due to early rejection). In case that one continues to the final stage, the overall analysis at the end of the trial combines the results of all stages. Bauer and Köhne (1994) proposed to combine the evidence from different stages via the product of the stagewise p-values. A result due to Fisher shows that the logarithm of the reciprocal of the square root of this product follows a χ^2-distribution. An alternative adaptive design approach has been proposed by Proschan and Hunsberger (1995) in which a trial may be extended to subsequent stages based on the conditional error calculated at an interim analysis. The conditional error function is thus the probability to reject the null hypothesis in the final analysis given the observed information at interim under the assumption that the null hypothesis holds.

Lehmacher and Wassmer (1999) first established a connection between adaptive designs and group sequential tests. They combine the evidence from different stages via the use of weighted inverse normal functions of the observed p-values. This approach was extended by Müller and Schäfer (2001) to allow for the number and timing of the interim analyses to determine the critical values used at each interim analyses via the use of spending functions as in group sequential trials. It can be shown that if the interim analyses occur at the pre-planned information fractions and the design is not adapted during the conduct of the trial, the resulting stopping boundaries are the same as in classical group sequential methods but not the test statistics (Kelly et al., 2005). In the same spirit, König et al. (2008) proposed an adaptive Dunnett test procedure based on the conditional error rate of the single stage Dunnett test. The adaptive procedure uniformly improves the classical Dunnett test, which is shown to be strictly conservative if treatments are dropped at interim.

Hommel (2001) provided a general framework for adaptively changing hypotheses at interim. This method is based on the closure principle (Marcus et al., 1976) for controlling the overall type I error rate. The closure principle considers all intersection hypotheses constructed from the initial hypotheses set of all pairwise comparisons with the control. A dose is declared significant, if all intersection hypotheses related to this dose are also rejected. Hommel (2001) proposed to apply the closure principle in the framework of adaptive tests by testing each intersection hypothesis with a suitable combination function to summarize the evidences across the different stages. In the meantime, a number of further considerations related to confirmatory adaptive designs with treatment or dose selection at interim have been investigated, see Bauer and Kieser (1999), Posch et al. (2005), Bretz et al. (2006) and Schmidli et al. (2006).

Bayesian adaptive dose finding designs are an important alternative to the confirmatory flexible designs discussed previously. Information can either be updated continuously as data is accrued in the trial, or in cohorts of patients. This makes this class of designs very appealing to sequential decision making and experimentation, including clinical studies. Bayesian approaches enable the calculation of predictive probabilities of future results for any particular design, which allows the comparison of designs on the basis of probabilities of their consequences. Although control of the type I error rate is not an intrinsic property of a Bayesian design, simulations can be used to tailor a Bayesian adaptive trial such that it maintains this and other desirable frequentist operational characteristics conditional on the dose selection or adaptation rules applied in the simulations. A potential downside to the Bayesian approach is the computational complexity coupled with the absence of commercial software packages to assist with study design and analysis. We refer to Berry et al. (2001) and Krams et al. (2003) for more methodological details and an example clinical study employing Bayesian dose finding methods in practice.

Alternatively, adaptive (Bayesian) optimal dose finding designs might be considered. Miller et al. (2007), for example, recognized the inherent model uncertainty and anticipated different dose response scenarios. Additionally, a-priori probabilities are assigned to the different scenarios, which reflect the prior believes of the clinical team. Optimal design theory is then applied to search numerically for the experimental design, which maximizes the mean efficiency (weighted by the prior probabilities) in comparison to a balanced design. Emerging information from interim looks can further be taken into account to modify the allocation ratio for the dose arms based on the Bayesian optimal design results.

Yet a different approach was proposed by Ivanova et al. (2008). They extended the method of Bretz et al. (2005) by adaptively allocating patients to improve the efficiency of the target dose estimation step. A modification of a cumulative cohort design is suggested, where a dose is repeated if the current estimated difference in response between the dose and placebo scaled by the variance is close to the target effect and changed otherwise. At each step the t statistic comparing the difference between the mean response at the current dose with the mean response at placebo (and accounting for the clinical relevance shift) is computed. For details on the updating algorithm and the proposed allocation rules we refer to Ivanova et al. (2008).

Finally, we come back to the PhRMA working group on "Adaptive Dose Ranging Studies", which was already mentioned in the Introduction. The working group has evaluated several existing adaptive and non-adaptive dose finding methods. The methods considered comprise a representative cross-section of currently available dose finding procedures, including traditional ANOVA-based methods using Dunnett test, the MCP-Mod methodology introduced in Section 5, the application of Bayesian model averaging techniques, non-parametric dose response modeling approaches using local quadratic regression techniques as well Bayesian model-based adaptive designs and D-optimal response-adaptive approaches. Through an extensive simulation study based on a common set of scenarios (sample sizes, number of doses, etc.) for all procedures, the strengths and weaknesses of each method were investigated, in particular with respect to the ability of the procedures to learn from the data and adapt to emerging information. We refer to the White Paper published by the working group for a detailed summary of the simulation study (Bornkamp et al., 2007).

7 Conclusions

Dose finding studies play a key role in any drug development program and are often the gate-keeper for the large confirmatory studies in Phase III. Many approaches exist for the proper design and analysis of these trials. The ultimate choice of the method to be applied depends on the particular settings and goals. Dose finding studies should thus be tailored to best fit the needs of the particular drug development program under consideration. Methods are available, for example, to allow the conduct of seamless proof-of-concept and dose finding studies. Alternatively, if it is desired to extend dose finding trials straight into a confirmatory Phase III study, adaptive designs offer efficient possibilities to control the overall type I error rate at a pre-specified level. In summary, we encourage the consideration and implementation of advanced dose finding methods, which efficiently make use of accumulating information during the drug development process.

Acknowledgements Jason Hsu's research was supported by the National Science Foundation (NSF); Grant Number: DMS-0505519))

Conflict of interests statement
The authors have declared no conflict of interest.

References

Bates, D. and Chambers, J. (1992). Nonlinear models. pp. 421–454 in Chambers, J. M. and Hastie, T. J. (eds.) *Statistical Models in S*. New York, Chapman & Hall.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. Wiley, New York.

Bauer P. and Köhne K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18**, 1833–1848

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. Pp. 99–181 in Gatsonis, C. Carlin, B., and Carriquiry, A. (eds.) *Case Studies in Bayesian Statistics V.* Springer, New York.

Bornkamp, B. (2006). Comparison of model-based and model-free approaches for the analysis of dose response studies. Diploma thesis, University of Dortmund.

Bornkamp, B., Bretz, F., Dmitrienko, A., Enas G., Gaydos, B., Hsu, C. H., König, F., Krams, M., Liu, Q., Neuenschwander, B., Parke, T., Pinheiro, J., Roy, A., Sax, R., and Shen, F. (2007). Innovative approaches for designing and analyzing adaptive dose-ranging trials (with discussion). *Journal of Biopharmaceutical Statistics* **17**(6), 965–995.

Bornkamp, B., Pinheiro, J., and Bretz, F. (2008). MCPMod – An R package for the design and analysis of dose-finding studies. (Submitted)

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

Bretz, F. (2006). An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics & Data Analysis* **50**, 1735–1748.

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

Bretz, F., Hothorn, T., and Westfall P. (2008). Multiple comparison procedures in linear models. In: *COMPSTAT 2008 – Proceedings in Computational Statistics*, Brito, P. (ed.). Springer, Heidelberg (in press).

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society Series A* **158**, 419–466.

Chevret S. (2006). *Statistical methods for dose finding experiments*. Wiley, New York.

Dette, H., Bretz, F., Pepelyshev, A., and Pinheiro, J. C. (2007). Optimal Designs for Dose Finding Studies. *Journal of the American Statistical Association* (in press).

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

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

FDA (2004). “Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products.” FDA report from March 2004, available at http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html

FDC Report (1991). FDC Reports from May 6, 1991. *The Pink Sheet* **53**(18), 14–15.

Finner, H. and Strassburger, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *Annals of Statistics* **30**, 1194–1213.

Freund, R. and Littell, R. (2000). *SAS System for regression*. Wiley, New York.

Genz, A. and Bretz, F. (2002). Methods for the computation of multivariate *t*-probabilities. *Journal of Computational and Graphical Statistics* **11**, 950–971.

Hjorth, J. S. U. (1994). *Computer intensive statistical methods – Validation, model selection and bootstrap*. Chapman & Hall, London.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple comparisons procedures*. Wiley, New York.

Holford, N. (2006). Dose response: Pharmacokinetic-pharmacokdynamic approach. pp. 73–88 in Ting, N. (ed.) *Dose finding in drug development*. Springer, New York.

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* **43**, 581–589.

Hothorn, LA (2006). Multiple comparisons and multiple contrasts in randomized dose response trials – confidence interval oriented approaches. *Journal of Biopharmaceutical Statistics* **16**, 711–731.

Hothorn, T., Bretz, F., and Genz, A. (2001). On multivariate t and Gauss probabilities in R. *R Newsletter* **1**(2), 27–29.

Hothorn, T., Bretz, F., and Westfall, P. H. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**, 346–363.

Hsu, J. C. (1996). *Multiple comparisons*. Chapman and Hall, New York.

Hsu, J. C. and Berger, R. L. (1996). Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies *Journal of the American Statistical Association* **94**, 468–482.

ICH-E4 (1994). ICH Harmonized Tripartite Guideline. Topic E4: Dose-response information to support drug registration. Available at http://www.emea.eu.int

ICH-E9 (1998). ICH Harmonized Tripartite Guideline. Topic E9: Statistical principles for clinical trials. Available at http://www.emea.eu.int

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.

Jennison, C. and Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman and Hall, London.

Kelly P. J., Sooriyarachchi M. R., Stallard N., and Todd S. (2005). A practical comparison of group-sequential and adaptive designs. *Journal of Biopharmaceutical Statistics* **15**, 719–738.

König, F., Brannath, W., Bretz, F., and Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* **27**, 1612–1625.

Krams M., Lees K. R., Hacke W., Grieve A. P., Orgogozo J. M., and Ford G. A. (2003). Acute stroke therapy by inhibition of neutrophils (ASTIN): An adaptive dose response study of UK-279,276 in acute ischemic stroke. *Stroke* **34**, 2543–2548.

Krishna R. (2006). *Dose optimization in drug development*. Informa Healthcare, New York.

Lehmacher W. and Wassmer G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–1290.

Letter, P. (2000). *Paracelsus*. Königsfurt Verlag, Klein Königsförde, Germany.

Loader, C. (1999). *Local regression and likelihood*. New York, Springer.

Marcus R., Peritz E., and Gabriel K. B. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

McCullagh, P. and Nelder, J. A. (2007). *Generalized linear models*. New York, Chapman & Hall.

Miller, F., Dette, H., and Guilbaud, O. (2007). Optimal designs for estimating the interesting part of a dose effect curve. *Journal of Biopharmaceutical Statistics* **17**(6), 1097–1115.

Müller H. H. and Schäfer H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886–891.

Neal, T. (2006). Hypothesis testing and Bayesian estimation using a sigmoid Emax model applied to sparse dose response designs. *Journal of Biopharmaceutical Statistics* **16**(5), 657–677.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York, Springer-Verlag.

Pinheiro, J., Bretz, F. and Branson, M. (2006a). Analysis of dose response studies: Modeling approaches. pp. 146–171 in Ting, N. (ed.) *Dose finding in drug development*. Springer, New York.

Pinheiro, J., Bornkamp, B., and Bretz, F. (2006b). Design and analysis of dose finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics* **16**(5), 639–656.

Posch, M., König, F., Branson, M., Dunger-Baldauf, C. and Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* **24**, 3697–3714.

Proschan M. A. and Hunsberger S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.

Radetsky, M. (1996). The discovery of Penicillin. *Pediatric Infection Disease Journal* **15**, 811–818.

Ratkowsky, D. A. (1989). *Handbook of nonlinear regression models*. Marcel Dekker, New York.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988) *Order restricted statistical inference*. Wiley, New York.

Ruberg, S. J. (1995a). Dose response studies I. Some design considerations. *Journal of Biopharmaceutical Statistics* **5**, 1–14.

Ruberg, S. J. (1995b). Dose response studies II. Analysis and interpretation. *Journal of Biopharmaceutical Statistics* **5**, 15–42.

Schmidli, H., Bretz, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Applications and practical considerations. *Biometrical Journal* **48**(4), 635–643.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear regression*. New York: Wiley.

Stefansson, G., Kim, W.-C., and Hsu, J. C. (1988). On confidence sets in multiple comparisons. In: *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **2**, 89–104. Academic Press, New York.

Sun. H. and Fadiran, E. O. (2007). Pharmacometrics Applications in Population Exposure-Response Data for New Drug Development and Evaluation, pp. 937–954, in *Pharmacometrics: The Science of Quantitative Pharmacology*, Ette, E. I. and Williams, P. J. (ed.), J. Wiley & Sons: New York, NJ

Ting N. (2006). *Dose finding in drug development*. Springer, New York.

Tukey, J. W., Ciminera, J. L., and Heyse, J. F. (1985). Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* **41**, 295–301.

Turner, H. and Firth, D. (2007). Generalized Nonlinear Models in R. Statistical *Computing & Graphics Newsletter* **18**(1) 11–16, Alexandria, VA: American Statistical Association.

Unkelbach, H. D. and Wolf, T. (1985). *Qualitative Dosis-Wirkungs-Analysen*. Gustav Fischer Verlag, Stuttgart.

Wakana, A., Yoshimura, I., and Hamada, C. (2007). A method for therapeutic dose selection in a phase II clinical trial using contrast statistics. *Statistics in Medicine* **26**(3), 498–511.