Biometrics 61, 738–748
September 2005

DOI: 10.1111/j.1541-0420.2005.00344.x

Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies

F. Bretz,[1,*] J. C. Pinheiro,[2] and M. Branson[1]
[1]Novartis Pharma AG, Lichtstrasse 35, Basel, Switzerland
[2]Novartis Pharmaceuticals, One Health Plaza, East Hanover, New Jersey 07936, U.S.A.
[*]*email:* frank.bretz@novartis.com

Summary. The analysis of data from dose-response studies has long been divided according to two major strategies: multiple comparison procedures and model-based approaches. Model-based approaches assume a functional relationship between the response and the dose, taken as a quantitative factor, according to a prespecified parametric model. The fitted model is then used to estimate an adequate dose to achieve a desired response but the validity of its conclusions will highly depend on the correct choice of the a priori unknown dose-response model. Multiple comparison procedures regard the dose as a qualitative factor and make very few, if any, assumptions about the underlying dose-response model. The primary goal is often to identify the minimum effective dose that is statistically significant and produces a relevant biological effect. One approach is to evaluate the significance of contrasts between different dose levels, while preserving the family-wise error rate. Such procedures are relatively robust but inference is confined to the selection of the target dose among the dose levels under investigation. We describe a unified strategy to the analysis of data from dose-response studies which combines multiple comparison and modeling techniques. We assume the existence of several candidate parametric models and use multiple comparison techniques to choose the one most likely to represent the true underlying dose-response curve, while preserving the family-wise error rate. The selected model is then used to provide inference on adequate doses.

Key words: Contrast test; Dose finding; Minimum effective dose; Multiple testing.

1. Introduction

Identifying the right dose is a key goal in the development of any medicinal drug. Its importance cannot be understated: selecting too high a dose can result in an unacceptable toxicity profile, while selecting a dose that is too low increases the likelihood that the compound provides insufficient evidence of effectiveness. A critical component of this entire *decision process* is the dose and/or dose regimen, selected for clinical development. This process could be better supported through a framework that facilitates the combination of confirming the existence of a drug effect and an estimating dose(s) that provide a particular therapeutic response. Searching for an adequate dose has historically been addressed using two approaches: multiple comparisons and modeling. The former regards the dose as a qualitative factor and generally makes few, if any, assumptions about the underlying dose-response relationship. The latter assumes a functional relationship between the response and dose, taken to be a quantitative factor, according to a prespecified model.

When using multiple comparisons, one approach is to evaluate the significance of contrasts between different doses while preserving the family-wise error rate (FWER) at some prespecified level α. While such an approach is generally robust, statistical inference is restricted to the set of doses under investigation. Under an assumed functional dose-response relationship, the model-based approach provides flexibility as the fitted model may be used to provide estimates of primary interest, for example, estimating a dose required to achieve a desired level of response. The validity of such inference, however, may highly depend on having chosen an appropriate model.

In this article, we discuss a unified strategy that amalgamates multiple comparison and model-based approaches in the analysis of dose-response data. The approach consists of two major steps. We start with a set of potential models for the description of the dose-response data. From this candidate set, we select the "best" model (if any), while controlling the FWER by the use of multiple comparison procedures similar in spirit to ideas expressed in Shimodaira (1998). Upon selection of a model, the target dose of interest can then be estimated efficiently. This general concept may be considered an extension of the ideas initially proposed by Tukey, Ciminera, and Heyse (1985). Buckland, Burham, and Augustin (1997) also provide a framework in which uncertainty due to model selection is formally incorporated, but without adhering to a strong error control.

This article is organized as follows. Basic concepts on multiple comparison procedures and modeling techniques are reviewed in Section 2. In Section 3, the integration of both strategies is discussed and a new method is derived which accurately estimates the target dose within the dose range under investigation. In Section 4, we analyze a phase II dose

Table 1
A selection of frequently used dose-response models

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

finding study using the newly developed approach. In Section 5, simulation results are then presented and concluding remarks are given in Section 6.

2. Modeling and Multiple Comparisons

2.1 *Dose-Response Models*

The general framework adopted here 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 $d_2, d_3, \ldots, d_k$ plus placebo d_1, for a total of k arms. For the purpose of dose estimation, we consider the one-way layout for the model specification

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

where $\boldsymbol{\theta}$ refers to the vector of model parameters, i to the dose group ($i = 1$ corresponds to placebo), and j to the patient within dose group i.

For the purpose of determining optimal contrasts for model selection (see Section 3.1), without loss of generality, one needs only to consider a *standardized* version f^0 of the dose-response model $f(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(d, \boldsymbol{\theta}^0)$. Hence it is sufficient to consider strategies for producing prior estimates for $\boldsymbol{\theta}^0$ for the model testing contrasts.

Table 1 lists a (nonexhaustive) selection of models frequently used to represent dose-response relationships, together with their respective standardized versions. The quadratic model described in the table is assumed to be "umbrella shaped" with $\beta_2 < 0$. A graphical display for each of these models is shown in Figure 1.

Prior estimates for the standardized model parameters, $\boldsymbol{\theta}^0$, are typically derived from initial knowledge (or guesses) of the expected percentage p^* of the maximum response associated with a given dose d^*. For example, in the $E_{\max}$ model, an initial estimate for the single parameter in the standardized model based on (d^*, p^*) is given by $\widehat{ED}_{50} = d^*(1 - p^*)/p^*$. Similar expressions for all models in Table 1 are given in Branson, Pinheiro, and Bretz (2003).

2.2 *Contrast Tests*

For the purpose of detecting an overall trend, we assume the linear model

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2), \quad (2)$$

where $\mu_i = f(d_i, \boldsymbol{\theta})$ are the unknown treatment means with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$. Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ denote the arithmetic mean of group i with $\bar{\boldsymbol{Y}}' = (\bar{Y}_1, \ldots, \bar{Y}_k)$. Let further $S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2/\nu$ denote the pooled variance estimator with $\nu = \sum_i n_i - k$ degrees of freedom.

We assume that we are given a set $\mathcal{M} = \{M_m, m = 1, \ldots, M\}$ of M candidate models. Each of these models is parameterized through a fixed mean vector $\boldsymbol{\mu}_m^0 = (\mu_{m1}^0, \ldots, \mu_{mk}^0)'$ as derived from the standardized model $f_m^0(d, \boldsymbol{\theta}_m) = \boldsymbol{\mu}_m^0$. Note that $\mathcal{M}$ may contain parameter specifications from different models $f_m^0 \neq f_{m'}^0$ or include different parameter specifications $\boldsymbol{\theta}_m \neq \boldsymbol{\theta}_{m'}$ for the same model $f_m^0 = f_{m'}^0$, $1 \leq m \neq m' \leq M$.

Note that equation (1) is based on the true (but unknown) functional model f. In contrast, equation (2) is used for hypothesis testing and selection, that is, it is used to select the model (out of the candidate set $\mathcal{M}$) which best describes the data under investigation. As elaborated in Section 3, once a single model has been selected, we propose to continue using equation (1) for the dose-estimation step based on that selected model.

Our goal is to select the best fitting model(s) out of the candidate set $\mathcal{M}$, while controlling the FWER. To this end, we test the null hypotheses $H_0^m : \boldsymbol{c}_m' \boldsymbol{\mu} = 0$ against the one-sided alternatives $H_1^m : \boldsymbol{c}_m' \boldsymbol{\mu} > 0$ for given $k \times 1$ contrast vectors $\boldsymbol{c}_m' = (c_{m1}, \ldots, c_{mk})$ of known constants subject to $\boldsymbol{c}_m' \mathbf{1} = 0$, $m = 1, \ldots, M$. This leads to the single contrast tests

$$T_m = \frac{\boldsymbol{c}_m' \bar{\boldsymbol{Y}}}{\sqrt{S^2 \sum_{i=1}^{k} c_{mi}^2 / n_i}}, \quad m = 1, \ldots, M. \quad (3)$$

Under the assumptions of model (2) and H_0^m, each test statistic T_m has a central t-distribution with ν degrees of freedom. When H_0^m is not true, T_m follows a noncentral t-distribution with noncentrality parameter $\tau_m = \boldsymbol{c}_m' \boldsymbol{\mu} / (\sigma^2 \sum_{i=1}^{k} c_{mi}^2 / n_i)^{\frac{1}{2}}$. Because we want to select the best model(s), our focus is on simultaneous inferences about the parameters $\boldsymbol{c}_m' \boldsymbol{\mu}$. One way to combine the test statistics T_m into a single decision rule is to take the best contrast, that is, to take the maximum $T_{\max} = \max_m T_m$. If $T_{\max} > q$ for an appropriate critical value q, a significant dose-response signal is established. If P_m denotes the associated (multiplicity) adjusted P-value of T_m, an equivalent decision is obtained through $\min_m P_m < \alpha$. Moreover, model $M_{m'}$, $m' = \operatorname{argmin}_m P_m$ is selected for further investigation, as well as any other model with individual adjusted P-value being less than α (see Section 3.2).

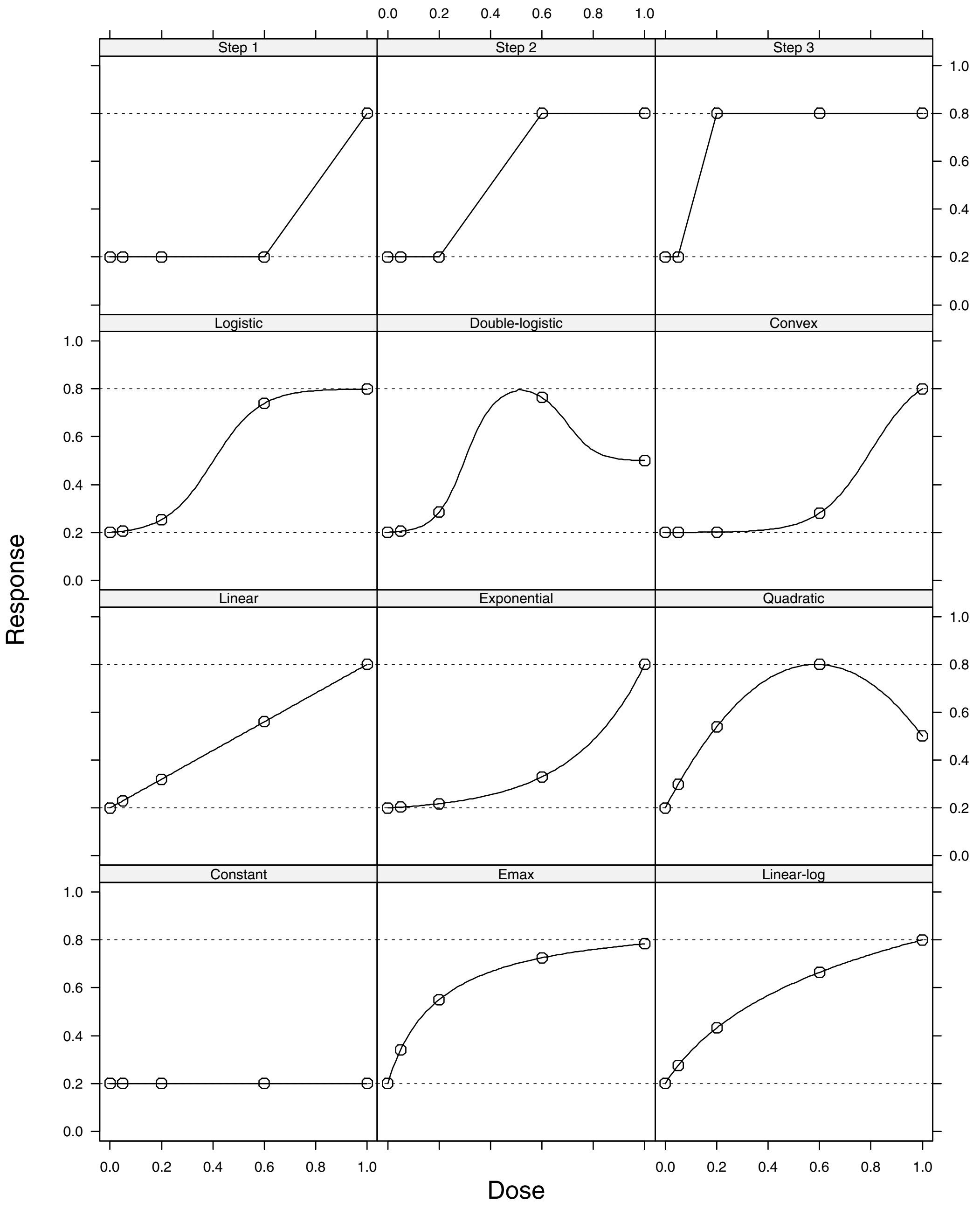

Figure 1. Dose-response profiles for shapes included in simulation study. Open dots indicate the responses at the dose levels used in the simulated study design.

The computation of the critical value q should account for the multiplicity to control the FWER at a prespecified level α (Hochberg and Tamhane, 1987). We thus consider the joint distribution of the vector $\boldsymbol{T}' = (T_1, \ldots, T_M)$. Under the assumptions above, $\boldsymbol{T} \sim MVT_M(\nu;\ \mathbf{0},\ \boldsymbol{R})$ is M-variate t-distributed with ν degrees of freedom and correlation matrix $\boldsymbol{R} = (\rho_{ij})$, where $\rho_{ij} = (\sum_{\ell=1}^{k} c_{i\ell}c_{j\ell}/n_\ell)/(\sum_{\ell=1}^{k} c_{i\ell}^2/n_\ell \sum_{\ell=1}^{k} c_{j\ell}^2/n_\ell)^{\frac{1}{2}}, 1 \leq i, j \leq k$. In the asymptotic case $\nu \to \infty$ or if σ is known, the corresponding limiting multivariate normal distribution holds. Numerical integration methods to compute the associated probabilities are described by Genz and Bretz (2002). Finally we note that the contrast tests above and the related results in this article can also be extended to general linear models, including covariates and factorial treatment structures.

3. Dose Finding Using Multiple Comparisons and Modeling

In this section we describe an approach which combines *m*ultiple *c*omparison *p*rocedures with *mod*eling techniques, abbreviated *MCP-Mod*. We assume the existence of several candidate parametric models and use multiple comparison techniques to choose the one most likely to represent the underlying dose-response curve. This is implemented in several steps. The first step consists of computing optimum contrast coefficients, once the doses, sample sizes, and the set of candidate models have been identified (Section 3.1). In the second step we evaluate the significance of the individual models in terms of the corresponding single contrast tests. Good candidate models are chosen which lead to significant contrast tests. Such a procedure allows the selection of the most adequate dose-response model, while preserving the FWER. Other criteria may be used for the model selection step, which is also discussed briefly (Section 3.2). Finally, the selected model is then used to produce inference on adequate doses, employing a model-based approach. Different decision rules for estimating the minimum effective dose (*MED*) are introduced in Section 3.3.

3.1 *Optimal Model Contrasts*

We consider the problem of determining the "best" contrast associated with a given model function $f(d, \boldsymbol{\theta})$, in the sense that, when that model is correct, it maximizes the chance of rejecting the associated null hypothesis, that is, it maximizes the noncentrality parameter $\tau = \tau(\boldsymbol{c})$. We first consider the two-sided case. Thus, one should choose $\boldsymbol{c}_{\text{opt}}(f)$ such that

$$\boldsymbol{c}_{\text{opt}}(f) = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, g(\boldsymbol{c}, \boldsymbol{\mu}),$$

$$g(\boldsymbol{c}, \boldsymbol{\mu}) = \frac{(\boldsymbol{c}'\boldsymbol{\mu})^2}{\displaystyle\sum_{i=1}^{k} c_i^2 \big/ n_i} = \sigma^2[\tau(\boldsymbol{c})]^2. \tag{4}$$

Such a $\boldsymbol{c}_{\text{opt}}(f)$ is only defined up to a scale factor, because $g(\boldsymbol{c}, \boldsymbol{\mu}) = g(\lambda\boldsymbol{c}, \boldsymbol{\mu})$, and so if $\boldsymbol{c}_{\text{opt}}$ maximizes $g(\boldsymbol{c}, \boldsymbol{\mu})$, then so does $\lambda\boldsymbol{c}_{\text{opt}}, \forall\lambda \in \mathbb{R}$. For uniqueness, we will require here that $\|\boldsymbol{c}_{\text{opt}}\| = 1$, where $\|\cdot\|$ refers to the L_2-norm. Note that such a $\boldsymbol{c}_{\text{opt}}$ is still not unique, because $\pm\boldsymbol{c}_{\text{opt}}$ are both optimal. These correspond to the optimal contrast vectors for the two possible one-sided tests. For two-sided tests, however, it is irrelevant which $\boldsymbol{c}_{\text{opt}}$ is chosen.

The two one-sided tests associated with H_0 have alternative hypotheses $H_1^+ : \boldsymbol{c}'\boldsymbol{\mu} > 0$ and $H_1^- : \boldsymbol{c}'\boldsymbol{\mu} < 0$. Let $\pm\boldsymbol{c}_{\text{opt}}^{2s}$ be the solutions corresponding to the two-sided test satisfying (4) and, without loss of generality, assume that $(\boldsymbol{c}_{\text{opt}}^{2s})'\boldsymbol{\mu} > 0$. It follows that $\boldsymbol{c}_{\text{opt}}^{2s}$ is the optimal contrast for the one-sided test corresponding to H_1^+ and $-\boldsymbol{c}_{\text{opt}}^{2s}$ is the optimal contrast corresponding to H_1^-. To see that, assume that there exists a $\boldsymbol{c}_{\text{opt}}^+ \neq \boldsymbol{c}_{\text{opt}}^{2s}$ that maximizes $\tau(\boldsymbol{c})$. This would imply that $\tau(\boldsymbol{c}_{\text{opt}}^+) > \tau(\boldsymbol{c}_{\text{opt}}^{2s})$, which, in turn, would give $g(\boldsymbol{c}_{\text{opt}}^+, \boldsymbol{\mu}) > g(\boldsymbol{c}_{\text{opt}}^{2s}, \boldsymbol{\mu})$, a contradiction, because $\boldsymbol{c}_{\text{opt}}^{2s}$ maximizes $g(., \boldsymbol{\mu})$. A similar argument applies to H_1^-.

Assuming that there exists a standardized version f^0 of f such that $\boldsymbol{\mu} = \boldsymbol{\mu}(f) = \theta_0 + \theta_1\boldsymbol{\mu}(f^0) = \theta_0 + \theta_1\boldsymbol{\mu}^0$, it is easy to verify that $\boldsymbol{c}'\boldsymbol{\mu} = \theta_1\boldsymbol{c}'\boldsymbol{\mu}^0$ and $g(\boldsymbol{c}, \boldsymbol{\mu}) = \theta_1^2\, g(\boldsymbol{c},\boldsymbol{\mu}^0)$. Therefore, $\boldsymbol{c}_{\text{opt}}(f) = \boldsymbol{c}_{\text{opt}}(f^0)$ and it suffices to consider the problem of finding the $\boldsymbol{c}_{\text{opt}}$ corresponding to the standardized model f^0.

We consider first the calculation of $\boldsymbol{c}_{\text{opt}}$ under balanced sample size allocation $n_i = n$, $i = 1, \ldots, k$, in which case a closed-form solution exists. In this case, under the restriction $\|\boldsymbol{c}\| = 1$, $g(\boldsymbol{c}, \boldsymbol{\mu}) = n(\boldsymbol{c}'\boldsymbol{\mu})^2$, it suffices to maximize $(\boldsymbol{c}'\boldsymbol{\mu})^2$ in $\boldsymbol{c}$. It follows from the Cauchy–Schwarz inequality and our assumptions on $\boldsymbol{c}$ that $(\boldsymbol{c}'\boldsymbol{\mu})^2 = [\boldsymbol{c}'(\boldsymbol{\mu} - \bar{\mu}\mathbf{1})]^2 \leq \|\boldsymbol{\mu} - \bar{\mu}\mathbf{1}\|^2$, where $\bar{\mu} = \boldsymbol{\mu}'\mathbf{1}/k$. Thus, it follows from the discussion above on the sufficiency of considering the standardized model that under balanced sample size allocation

$$\boldsymbol{c}_{\text{opt}} = \frac{\boldsymbol{\mu} - \bar{\mu}\mathbf{1}}{\|\boldsymbol{\mu} - \bar{\mu}\mathbf{1}\|} = \frac{\boldsymbol{\mu}^0 - \bar{\mu}^0\mathbf{1}}{\|\boldsymbol{\mu}^0 - \bar{\mu}^0\mathbf{1}\|}.$$

In the more general case of unequal sample sizes per treatment arm, $\boldsymbol{c}_{\text{opt}}$ cannot be expressed in closed form and numerical optimization techniques are required. By assumption, $\boldsymbol{c}'\mathbf{1} = 0$ and $\|\boldsymbol{c}\| = 1$, so that $\boldsymbol{c}$ can be expressed as a function of $k - 2$ free parameters $\boldsymbol{\delta}$, that is, $\boldsymbol{c} = h(\boldsymbol{\delta})$ for some parameterization function h. For numerical optimization purposes, it is often easier, more stable, and more robust to use a parameterization h for which the elements of $\boldsymbol{\delta}$ are unconstrained. With such an unconstrained parameterization, general purpose optimization software can be used to obtain $\boldsymbol{c}_{\text{opt}}$ via maximization of the objective function $g(h(\boldsymbol{\delta}), \boldsymbol{\mu})$, g defined as in (4). In the following we describe one possible choice for h, based on a spherical parameterization (Pinheiro and Bates, 1996).

Because $\boldsymbol{c}$ takes values on the surface of the unit sphere in $\mathbb{R}^k$, its elements can be expressed in spherical coordinates as

$$c_i = \sin(\gamma_i)\prod_{j<i}\cos(\gamma_j), \quad i = 1, \ldots, k-1,$$

$$c_k = \prod_{j<k}\cos(\gamma_j), \tag{5}$$

for a set of angles $\gamma_1, \ldots, \gamma_{k-1} \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Note that we have reduced the number of parameters by 1 because the condition $\|\boldsymbol{c}\| = 1$ is implicitly satisfied in (5). The second restriction $\boldsymbol{c}'\mathbf{1} = 0$ is established by noting that $c_1 = \sin(\gamma_1) = -\sum_{i>1} c_i = -\cos(\gamma_1)\sum_{i>1} c_i^1$, where $c_i^1 = c_i/\cos(\gamma_1)$. It then follows that $\gamma_1 = \tan^{-1}(-\sum_{i>1} c_i^1)$, so that we only require the angles $\gamma_2, \ldots, \gamma_{k-1}$. These are free, but constrained to the box $(-\frac{\pi}{2}, \frac{\pi}{2})^{k-2}$. We obtain an unconstrained parameterization by defining $\delta_i = \log(\pi/2 + \gamma_i/\pi/2 - \gamma_i)$ or, equivalently, $\gamma_i = -\pi/2 + \pi/(1 + \exp(-\delta_i))$. This parameterization can be used to obtain the optimum contrast $\boldsymbol{c}_{opt}$ using standard optimization software. S code to compute $\boldsymbol{c}_{opt}$ for given vectors $\boldsymbol{\mu}$ and $\boldsymbol{n} = (n_1, \ldots, n_k)$ of means and sample sizes is available from the authors upon request.

3.2 *Model Selection*

Once we have optimal contrast coefficients for each model in $\mathcal{M}$, we evaluate the significance of the individual models in terms of the corresponding single contrast tests T_m. Each T_m translates into a decision procedure, whether a selected dose-response curve $f_m(d, \boldsymbol{\theta})$ is significant given the current data, while controlling the FWER at level α. To maintain the nominal size α, the definition of $\boldsymbol{\mu}_m^0 = \boldsymbol{\mu}(f_m^0)$ is required prior to the experiment. A good candidate curve is chosen as one with a significant contrast test, given that its individual multiplicity adjusted P-value is less than α. If we do not obtain a significant contrast test, no model is selected from the set $\mathcal{M}$, and the modeling component of the *MCP-Mod* procedure is not undertaken. Such a result does not necessarily imply that the compound under investigation has no effect. Possible reasons for statistical nonsignificance could include small sample sizes or high variance. In addition, the set $\mathcal{M}$ might have been poorly chosen, such that the candidate models do not fit the true curve.

If at least one model is significant, a *reference set* of good models is obtained. Each model in the reference set is then statistically significant at FWER α and approximates the true model satisfactorily, as it follows from the diagnostic property of maximum tests described by Cox (1977). In addition, the true model should be associated with the minimum P-value by construction of the *MCP-Mod* method: In Section 3.1 we maximized the noncentrality parameter for the optimum choice of the contrast coefficients, which is equivalent to maximizing the correlation between $\boldsymbol{c}_m$ and $\boldsymbol{\mu}_m^0$. If the true dose-response shape coincides with $\boldsymbol{\mu}_m^0$ for some m, T_m would be the most powerful test among all contrast tests to detect this particular dose-response shape (Abelson and Tukey, 1963). These considerations give rise to concerns on the reliability of $\boldsymbol{\mu}_m^0$, which depends on the prior estimates $\boldsymbol{\theta}_m$ associated with f_m^0. In our experience we have not seen this issue to be of major concern and *MCP-Mod* tends to behave robustly against moderate misspecifications of $\boldsymbol{\theta}$. In cases where prior knowledge about $\boldsymbol{\theta}$ is scarce or not available, one may include different parameter specifications for the same model in the initial candidate set (see Section 2.2). In contrast to a direct application of a model-based approach, the preliminary steps of our approach address issues of possible model misspecifications and include the associated statistical uncertainty in a rigorous hypothesis testing framework.

In practice, it might be hard to decide upon the best model. If the second-best model has a P-value lying close to the minimum P, both models might be worth later consideration and additional data acquisition or further decision elements might be required. In other cases, fitting the model with the highest T-value (or equivalently, min P-value) is not possible because of numerical instabilities. Such problems occur, for example, when the model to be fitted contains many parameters in comparison to the number of doses, or when the doses are not spread appropriately throughout the dose range under investigation. The recommended procedure is to progress through the "ordered" reference set until model convergence is achieved or the reference set is exhausted. An alternative is to order the reference set by an appropriate information criterion. Such information criteria, as the Akaike Information Criterion for example, typically involve the log-likelihood and an associated penalty term for the number of model parameters thus ensuring that, once a model is selected, a numerically feasible fit to the data is provided. Details on fitting the models and estimating the parameters are given in Bates and Watts (1988) and Pinheiro and Bates (2000).

3.3 *Dose Selection*

Once the overall dose-response relationship has been established (proof of activity; PoA) and an adequate model has been selected, the final step of *MCP-Mod* consists of fitting the selected model to the data and estimating adequately the target dose(s) of interest. In the following discussion we restrict our attention to the estimation of the *MED*, although the ideas are equally applicable when estimating other target doses. Following Ruberg (1995), the *MED* is defined as the smallest dose which shows a clinically relevant and a statistically significant effect. Let Δ denote the clinically relevant difference, that is, the smallest relevant difference, by which we expect a dose to be better than placebo.

Two definitions of the *MED* are possible, depending on whether the target dose is selected out of the discrete dose set $\mathcal{D} = \{d_1, \ldots, d_k\}$ under investigation or from the entire dose range $(d_1, d_k]$. In the former case, $MED = \text{argmin}_{d_i \in \mathcal{D}}\{\mu_i > \mu_1 + \Delta\}$. Typically, estimation of $MED \in \mathcal{D}$ in this framework is conducted by applying appropriate multiple testing procedures; see Tamhane, Dunnett, and Hochberg (1996). In contrast, model-based approaches allow $MED \in (d_1, d_k]$. Given a model $f(., \boldsymbol{\theta})$,

$$MED = \underset{d \in (d_1, d_k]}{\text{argmin}}\{f(d, \boldsymbol{\theta}) > f(d_1, \boldsymbol{\theta}) + \Delta\}. \qquad (6)$$

Note that 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.

In the following we focus on definition (6) and propose three different rules for estimating the true *MED*. Denote by L_d (U_d) the lower (upper) $1 - 2\gamma$ confidence limit of the predicted mean value p_d at dose d based on the model $f(., \boldsymbol{\theta})$ (as computed, for example, by the *nls* function in the `S-PLUS` and `R` languages; see Bates and Chambers, 1992). Thus, the choice of γ is not driven by the purpose of controlling type I error rates, in contrast to the selection of α for controlling the FWER for the PoA. The following alternative estimates are investigated further in Section 5:

$$\widehat{MED}_1 = \underset{d \in (d_1, d_k]}{\text{argmin}}\{U_d > p_{d_1} + \Delta, L_d > p_{d_1}\},$$

$$\widehat{MED}_2 = \underset{d \in (d_1, d_k]}{\text{argmin}}\{p_d > p_{d_1} + \Delta, L_d > p_{d_1}\},$$

$$\widehat{MED}_3 = \underset{d \in (d_1, d_k]}{\text{argmin}}\{L_d > p_{d_1} + \Delta\}.$$

By construction, $\widehat{MED}_1 \leq \widehat{MED}_2 \leq \widehat{MED}_3$ and it is seen from Section 5 that $\widehat{MED}_2$ tends to be less biased in estimating the true *MED* than the alternative estimates. Note that a dose obtained through any of the criteria above may not have a significant effect at level α, especially when γ is not small enough. Because γ and Δ are prespecified, it may happen that an *MED* is obtained which is lower than a dose in the study that had no significant effect. To avoid such problems, γ should be set reasonably small, perhaps even taking multiplicity due to the construction of the confidence bands into account (Scheffé, 1953). An alternative approach to guarantee statistical significance would be to use the confidence bounds L_{d-d_1} for the difference between the response at dose d and placebo d_1. One would then require $L_{d-d_1} > 0$ instead of $L_d > p_{d_1}$ in the estimates above for *MED*.

4. A Phase II Dose Finding Study

This was a randomized double-blind parallel group trial with a total of 100 patients being allocated to either placebo or one of four active doses coded as 0.05, 0.20, 0.60, and 1, with $n = 20$ per group. To maintain confidentiality, the actual doses have been scaled to lie within the [0, 1] interval. The response variable was assumed to be normally distributed and larger values indicate a better outcome. A priori the assumption of monotonicity $\mu_0 \leq \mu_{0.05} \leq \mu_{0.2} \leq \mu_{0.6} \leq \mu_1$ was made, where

Table 2
Pairwise comparisons to placebo for an example dose finding study

Parameter	Estimate	Standard error	t-value	1-sided P-value	Marginal 90% conf. int.
$\mu_1 - \mu_0$	0.6038	0.2253	2.68	0.0044	(0.2296, 0.9780)
$\mu_{0.6} - \mu_0$	0.5895	0.2253	2.62	0.0052	(0.2153, 0.9637)
$\mu_{0.2} - \mu_0$	0.4654	0.2253	2.07	0.0201	(0.0912, 0.8396)
$\mu_{0.05} - \mu_0$	0.1118	0.2253	0.50	0.3103	(−0.2623, 0.4860)

μ_d denotes the true response for dose d. A fixed sequence test (Westfall and Krishen, 2001) was adopted to determine efficacious doses using a 5% one-sided level. In Table 2, the pairwise comparisons of treatment to placebo based on a one-way analysis of variance model are summarized. As seen from Table 2, the fixed sequence test stopped after having concluded that the top three doses of treatment were statistically significantly different to placebo.

We re-analyze these data using the *MCP-Mod* approach. The set of candidate models includes $E_{\max}$, linear in log-dose, linear, two exponential, and two quadratic (umbrella shape). The two variants of both the exponential and quadratic are included to ensure that we cover a broad dose-response ("shape space") within this initial candidate set. For the linear and linear in log-dose models, the values of the doses determine the optimal contrasts. For the five remaining models, additional parameters require specification. Methods for obtaining initial estimates based on the corresponding standardized models are discussed in Branson et al. (2003). In the case of the $E_{\max}$ model we require an estimate of the ED_{50}, that is, the dose providing half of the maximum change. For the purpose of illustration, we assume this value is 0.20. The contrast coefficients for the first exponential and the first quadratic models are also obtained by using the value for ED_{50} specified above in combination with the methods described in Branson et al. (2003). For the the second quadratic, we define the maximum to occur at dose 0.5, while for the second exponential, we define the parameter δ to be 0.15—this corresponds to a dose-response for which there is a negligible dose effect for the first three active doses and then a sharp rise in the expected value of the response.

A summary of the multiple comparison component of *MCP-Mod*, ordered by the magnitude of the observed t-values, is given in Table 3. The one-sided unadjusted (raw) P-values are presented and are accompanied with the corresponding adjusted one-sided P-values from the multivariate t-distribution. All contrast tests with an adjusted P-value less than 0.05 or equivalently with a t-value greater than 2.18 can be declared statistically significant having maintained the FWER at level 5%. Of the the seven candidate models, two model shapes are not statistically significant—the second exponential and the second quadratic. These two models do not therefore enter the reference set. It is also interesting to note that the first exponential model achieves borderline significance. These results are in accord with intuition upon inspection of the observed "dose response" summarized in Table 2.

Table 3
Summary of contrast tests

Contrast	Estimate	t-value	Raw P-value	Adjusted P-value
$E_{\max}$	0.552	3.46	0.0004	0.0017
Linear-log	0.524	3.29	0.0007	0.0028
Quadratic (1)	0.494	3.10	0.0013	0.0048
Linear	0.473	2.97	0.0019	0.0069
Exponential (1)	0.353	2.22	0.0145	0.0448
Exponential (2)	0.302	1.90	0.0304	0.0866
Quadratic (2)	0.295	1.85	0.0337	0.0950

The $E_{\max}$ model is the one chosen for the dose-selection component of this analysis. The parameter estimates (standard errors) were 0.321 (0.152), 0.746 (0.236), and 0.142 (0.180) for E_0, $E_{\max}$, and ED_{50}, respectively. Suppose that we had intended to use $\widehat{MED}_2$ with $\gamma = 0.05$, and we set the clinical relevance threshold $\Delta = 0.4$, that is, we wish to determine the smallest dose such that the lower limit of the 90% confidence interval for the predicted response is greater than the predicted response for placebo and that the point estimate is 0.4 above the predicted placebo level. Based on these assumptions, the selected dose is 0.17. Let us further assume that the set of active-dose levels used in this study defines the only doses for which it is possible to manufacture the experimental treatment. In this case, a pragmatic approach could then be to suggest that the implementation of the *MCP-Mod* approach leads to dose 0.2 being determined as the minimum dose that attains the prescribed conditions for selection. In principle, any dose lying above 0.17 may be defined as an acceptable dose, provided that the gain in efficacy does not result in an unacceptable safety risk.

5. Simulation Results

In this section we investigate, via simulation, the performance of the *MCP-Mod* approach. Two main aspects will be studied here: (i) the power to detect a dose-response relationship (PoA) and (ii) estimation of a dose close to the desired level taking into account both statistical significance and clinical relevance (*dose-selection* performance). Other classical dose-response tests were also included in the simulations for comparison with *MCP-Mod*, with respect to the PoA. Because model-based dose-selection methods can choose any value on a continuous scale, they are not directly comparable to dose finding methods based on multiple comparisons alone. The latter are restricted to selecting a dose from the set of doses

Table 4
Data generating dose-response shapes, $\mathcal{D} = \{0, 0.05, 0.2, 0.6, 1\}$

Model	Specification of $\mu(d)$
Constant	0.2
$E_{\max}$	$0.2 + 0.7d/(0.2 + d)$
Linear in log-dose	$0.2 + 0.6 \log(5d + 1)/\log(6)$
Linear	$0.2 + 0.6d$
Exponential	$0.183 + 0.017 \exp[2d \log(6)]$
Quadratic	$0.2 + 2.049d - 1.749d^2$
Logistic	$0.193 + 0.607/\{1 + \exp[10 \log(3)(0.4 - d)]\}$
Double-logistic	$\{0.198 + \frac{0.61}{1+\exp[18(0.3-d)]}\}I(d \leq 0.5) + \{0.499 + \frac{0.309}{1+\exp[18(d-0.7)]}\}I(d > 0.5)$
Truncated-logistic	$0.2 + 0.682/\{1 + \exp[10(0.8 - d)]\}$
Step 1	$0.2 + 0.6I(d \geq 1), d \in \mathcal{D}$
Step 2	$0.2 + 0.6I(d \geq 0.6), d \in \mathcal{D}$
Step 3	$0.2 + 0.6I(d \geq 0.2), d \in \mathcal{D}$

under investigation. Thus, the dose-selection performance is only investigated for *MCP-Mod*.

5.1 *Design*

The study design used for the simulations was based on that of the case study of Section 4. We investigated five dose levels ($d = 0$, 0.05, 0.2, 0.6, 1), with a single endpoint measured per patient, $Y \overset{\text{ind}}{\sim} \mathcal{N}(\mu(d), \sigma^2)$. Sample sizes per group were set to $n = 10$, 25, 50, 75, 100, and 150. The one-sided significance level for PoA was set at $\alpha = 0.05$. Twelve different data generating dose-response shapes for the mean responses $\mu(d)$ were investigated in the simulations as defined in Table 4. All of these shapes have the property that at $d = 0$ the response value is about 0.2 and, with the exception of the constant shape, all have a maximum response of about 0.8 within the interval [0, 1] (i.e., a maximum dose effect of about 0.6). Figure 1 displays the dose-response profiles for the 12 shapes listed in Table 4. A total of 10,000 simulated trials were generated for each shape × sample size combination.

The constant shape is included to evaluate the performance of the *MCP-Mod* method in terms of preserving the FWER for PoA determination. The second through seventh shapes in Table 4 have been described in Section 2.1 as typical dose-response models used in practice. They will form the set of candidate models for the contrast tests. Shapes eight and nine are included to evaluate the performance of the *MCP-Mod* method under model misspecification: They do not quite correspond to any of the models in the candidate set, though can be approximated by some of them. The last three shapes lead to a similar model misspecification, being the shapes for which the competing step contrasts introduced below are most powerful.

The test contrasts used in the simulations for the six models in the candidate set ($E_{\max}$, linear-log, linear, exponential,

Table 5
PoA probabilities for the different dose finding methods, under the shape × sample size combinations

		Data generating shape											
Methods	n	Const	$E_{\max}$	Lin-Log	Lin	Exp	Quad	Logist	D.Logist	T.Logist	Step 1	Step 2	Step 3
MCP-Mod	10	0.048	0.238	0.254	0.250	0.222	0.200	0.308	0.212	0.224	0.214	0.334	0.309
	25	0.055	0.462	0.481	0.478	0.444	0.371	0.599	0.399	0.436	0.419	0.636	0.595
	50	0.051	0.715	0.742	0.738	0.705	0.618	0.852	0.641	0.711	0.699	0.894	0.860
	75	0.053	0.853	0.883	0.875	0.851	0.785	0.951	0.796	0.855	0.859	0.973	0.954
	100	0.049	0.936	0.950	0.942	0.934	0.890	0.986	0.903	0.941	0.938	0.994	0.987
	150	0.049	0.986	0.990	0.989	0.990	0.974	0.999	0.976	0.990	0.990	1.000	0.999
LRT	10	0.046	0.241	0.250	0.246	0.221	0.172	0.285	0.181	0.224	0.219	0.303	0.310
	25	0.054	0.461	0.473	0.465	0.442	0.303	0.567	0.334	0.434	0.426	0.596	0.604
	50	0.051	0.714	0.730	0.721	0.695	0.519	0.825	0.549	0.705	0.702	0.873	0.877
	75	0.051	0.850	0.873	0.865	0.843	0.685	0.940	0.715	0.850	0.862	0.964	0.964
	100	0.048	0.935	0.945	0.935	0.930	0.800	0.981	0.845	0.936	0.941	0.991	0.991
	150	0.047	0.986	0.990	0.988	0.988	0.927	0.998	0.948	0.989	0.991	1.000	0.999
Step contrasts	10	0.048	0.226	0.234	0.232	0.210	0.169	0.273	0.178	0.216	0.212	0.297	0.302
	25	0.053	0.438	0.448	0.448	0.429	0.302	0.554	0.336	0.424	0.416	0.591	0.598
	50	0.052	0.692	0.707	0.702	0.683	0.518	0.818	0.553	0.696	0.697	0.872	0.876
	75	0.050	0.831	0.855	0.851	0.834	0.685	0.937	0.719	0.846	0.861	0.964	0.964
	100	0.048	0.923	0.933	0.925	0.924	0.796	0.980	0.845	0.933	0.940	0.992	0.991
	150	0.046	0.984	0.987	0.985	0.988	0.926	0.998	0.950	0.988	0.991	1.000	0.999

Table 6

Probability of correctly identifying the response model in the contrast testing step, for the six models in the candidate set and the different sample sizes, under $\sigma = 0.65$. Given in italics below is the associated probability of using the correct dose-response model.

n	$E_{\max}$	Lin-log	Linear	Exp	Quad	Logist
10	0.27	0.09	0.14	0.47	0.44	0.41
	0.39	*0.20*	*0.34*	*0.44*	*0.74*	*0.08*
25	0.51	0.21	0.29	0.76	0.78	0.64
	0.53	*0.28*	*0.41*	*0.57*	*0.84*	*0.15*
50	0.65	0.35	0.46	0.86	0.91	0.76
	0.65	*0.39*	*0.52*	*0.72*	*0.92*	*0.27*
75	0.73	0.45	0.57	0.90	0.95	0.82
	0.73	*0.47*	*0.60*	*0.82*	*0.96*	*0.35*
100	0.77	0.53	0.66	0.93	0.98	0.87
	0.77	*0.54*	*0.69*	*0.88*	*0.98*	*0.44*
150	0.84	0.65	0.76	0.97	0.99	0.93
	0.84	*0.66*	*0.77*	*0.95*	*0.99*	*0.55*

quadratic, and logistic) were obtained using the true parameter values for the corresponding standardized model (e.g., $ED_{50} = 0.2$ for the $E_{\max}$ model). When investigating the correlation of the test contrasts (not shown here), it transpires that there is considerable correlation between some of the models, indicating that it may be hard to discriminate between them in the simulations. The smallest correlation between any two of the six models is 0.26, while the linear, linear in log, and the $E_{\max}$ models, for example, have pairwise correlations above 0.9.

5.2 *PoA Performance*

For the purpose of evaluating the PoA performance of the *MCP-Mod* method, the response standard deviation (SD) was set at $\sigma = 1.478$, which, for a group sample size of $n = 75$, gives a power of 80% for the pairwise test between two doses at the maximum effect of $\delta = 0.6$. Table 5 gives the simulated probabilities of establishing PoA for the different methods under the various shape × sample size combinations.

We included the likelihood ratio test (LRT; Robertson, Wright, and Dykstra, 1988) and the step contrasts (Bauer and Hackl, 1985) as competitors to the *MCP-Mod*. The LRT is known to be one of the most powerful tests for trend throughout the order restricted alternative region $\mu_1 \leq \cdots \leq \mu_k$. In contrast to *MCP-Mod*, the LRT is designed for the PoA only and thus does not give any information about the underlying dose-response shape. Extensive mathematical details are given in Robertson et al. (1988). The step contrasts are a powerful alternative to the LRT. It can be shown that the step contrasts match exactly the corner vectors of the polyhedral cone described through the relationship $\mu_1 \leq \cdots \leq \mu_k$. Thus, they span the order restricted space of interest. The step contrasts are particularly powerful for finding the change point in a series of treatment means (Bauer and Hackl, 1985). For our simulations we used the multivariate t-distribution to compute the critical values (Genz and Bretz, 2002).

The FWER is well controlled at the 5% level for all sample sizes. Two of the three step shapes (Step 2 and Step 3) lead to the highest PoA power and the nonmonotone shapes (quadratic and double logistic) to the smallest. The PoA power values for the other shapes are of comparable magnitude. Note that *MCP-Mod* has power comparable to the LRT for all monotone shapes. *MCP-Mod* is considerably more powerful than the LRT for the quadratic and the double-logistic shapes, because the LRT is not designed for such downturns at higher doses. Both the *MCP-Mod* and the LRT are on average more powerful than the step contrasts. Note that even for the step shapes, for which we expect the step contrasts to be particularly powerful, *MCP-Mod* behaves similar to, if not better than, its competitors.

Table 7

Median relative bias and relative IQR of MED estimators under various shapes

		Median relative bias (%)			Relative IQR (%)		
Model	n	$\widehat{MED}_1$	$\widehat{MED}_2$	$\widehat{MED}_3$	$\widehat{MED}_1$	$\widehat{MED}_2$	$\widehat{MED}_3$
$E_{\max}$	50	−36.3	5.0	38.7	82.5	93.7	108.7
	150	−25.0	1.2	31.2	48.7	60.0	75.0
Linear in log-dose	50	−24.0	−0.1	23.8	56.5	63.0	71.7
	150	−13.1	−0.1	17.3	28.2	36.9	43.4
Linear	50	−22.0	−7.0	9.5	28.5	33.0	34.5
	150	−11.5	−2.5	8.0	15.0	18.0	19.5
Exponential	50	−17.0	−2.2	4.7	21.6	19.3	14.8
	150	−6.7	0.1	4.7	12.5	11.4	10.2
Quadratic	50	−23.3	1.0	25.2	24.2	36.4	48.5
	150	−11.1	1.0	21.2	16.2	20.2	32.3
Logistic	50	−11.8	5.4	29.0	38.7	40.9	45.2
	150	−7.5	3.2	20.4	30.1	30.1	32.3
Double-logistic	50	−28.7	−4.9	18.8	44.6	59.4	65.4
	150	−19.8	−4.9	12.9	23.8	32.7	41.6
Truncated-logistic	50	−9.0	5.4	12.6	20.4	18.0	10.8
	150	1.8	9.0	13.8	8.4	9.6	7.2

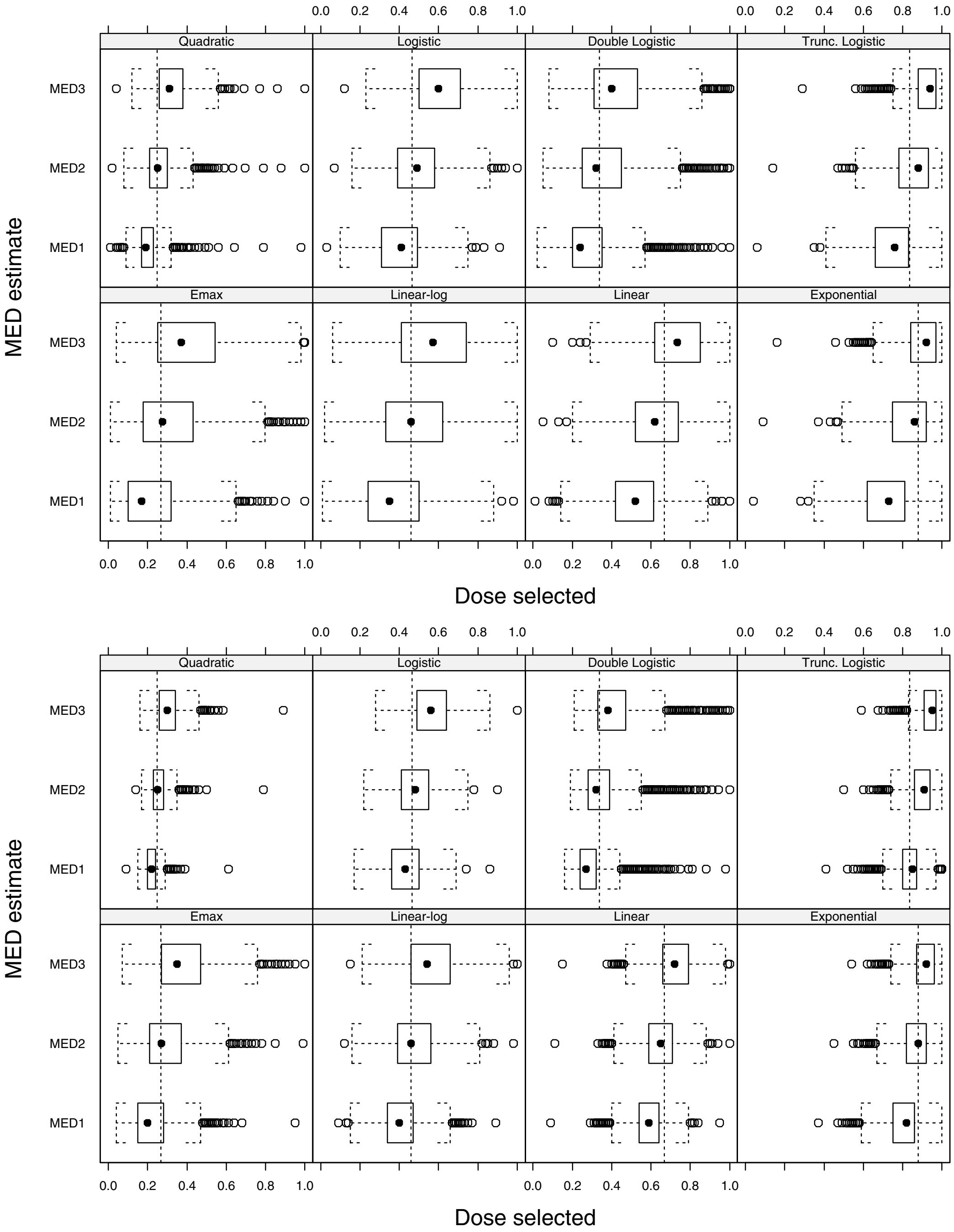

Figure 2. Boxplots of simulated dose selections for different MED estimators corresponding to $\gamma = 0.1$ with $n = 50$ (top) and $n = 150$ (bottom). Vertical dashed lines indicate target doses.

5.3 *Dose-Selection Performance*

For the purpose of evaluating the dose-selection performance of the *MCP-Mod* method, a response SD of $\sigma = 0.65$ was used. This is consistent with the estimated residual SD observed for the case study of Section 4 and provides a more realistic value for the simulations. As before, we used a clinically relevant effect of $\Delta = 0.4$. For the remainder of this section we consider only the second through ninth shape in Table 4, because the other shapes were only included for PoA comparison purposes.

Due to the smaller response SD, the simulated PoA power values were close to 1 for $n \geq 25$ and any dose-response shape, so that dose response is detected almost for sure. It is also useful to look at the probabilities of selecting the correct model in the contrast testing step, as well as the probabilities of using the correct model for the dose-selection step (which do not need to be the same because of potential convergence problems). Table 6 gives the corresponding estimated probabilities.

The exponential and the quadratic model have the best discrimination power. This is not surprising because their associated contrasts are the least correlated with the remaining model contrasts. The linear-log and linear model are the hardest to identify, which is again consistent with their high correlation. Because models which can represent similar dose-response profiles will likely lead to similar dose selections in the second stage of the method, the discrimination among highly correlated models is less critical than among the less correlated ones. For all but the exponential and the logistic model, there is an increase in the probability of choosing the correct model compared with that of selecting it based on the contrast tests. The basic problem with, for example, the logistic model in the simulation scenarios used here is that doses were not chosen to well characterize the shape of the logistic curve. Because there were only five doses (two of them, 0 and 0.05, fairly close to each other) and four parameters to be estimated in the logistic model, the unfavorable design led to frequent convergence problems when the logistic model was selected by the contrast testing step. A similar explanation holds also for the exponential model.

We now discuss the simulation results with respect to the dose-selection performance of the *MED* estimators, measured in terms of its proximity to and dispersion around the target dose (the doses producing an improvement of $\Delta = 0.4$ over placebo). Because the target doses differ with each model, the performance of the estimator $\widehat{MED}_i$ is measured in terms of its relative deviation R_i from the target dose, where $R_i = 100(\widehat{MED}_i - MED)/MED$. The median and interquartile range (IQR) of R_i in the 10,000 simulated dose selections then characterize the relative bias and variability of $\widehat{MED}_i$. We present the results for $n = 50$ and 150 and $\gamma = 0.1$ (corresponding to 80% level confidence intervals). Table 7 gives the corresponding summary statistics for various shapes. A graphical view of the dose-selection performance of the *MED* estimators is given by the boxplots of $\widehat{MED}_1$, $\widehat{MED}_2$, and $\widehat{MED}_3$ for the simulated trials, under the different shapes; see Figure 2.

It is clear from the table and the figures that $\widehat{MED}_1$ tends to underestimate the target dose, $\widehat{MED}_3$ tends to overestimate it, and $\widehat{MED}_2$ estimates the target dose more consistently. The precision of the methods is considerably enhanced when the sample size increases from 50 to 150. It should be noted, though, that the dose design used in the simulations was intended for a multiple comparison determination of the *MED*, and not for modeling. A more suitable choice of doses with modeling in mind would yield considerably better results, for the same overall sample sizes. This is a topic of future research. The precision of the dose-selection algorithms vary considerably with the underlying dose-response shape. The quadratic, truncated logistic and exponential shapes tend to lead to greater precision (for the particular scenarios used here, but not in general), while the remaining shapes give similar dispersions for the dose estimates. Similar results are also observed for other sample sizes and values of γ. We conclude that $\widehat{MED}_2$ seems preferable to the other methods.

6. Conclusions

We have described a combined strategy for analyzing dose finding studies, including a PoA assessment and a dose-selection step. PoA is tested in a first stage, using multiple comparison methods to identify statistically significant contrasts corresponding to a set of candidate models. If PoA has been established, the best model is then used for dose selection in subsequent stages.

The proposed method, termed *MCP-Mod*, is seen to maintain the FWER at its nominal level, has PoA power comparable to the LRT under monotone dose-response settings, and is better than the LRT under nonmonotone scenarios, provided the set of candidate models is broad enough. The clear advantage of this new approach, in comparison to more traditional multiple comparison dose finding methods, is its added flexibility in searching for and identifying an adequate dose for future drug development.

A number of issues remain topics of future research. Extensions of *MCP-Mod* to repeated measures (e.g., crossover designs, titration studies, etc.) require more investigation, including extensions to mixed-effects models. Other research questions include the investigation of heteroscedasticity across the models in equation (1), the sensitivity of the methods to the choice of initial values for the standardized models, the investigation of designs more suitable for dose-response modeling, and the development of statistical software implementing the methods and procedures described herein.

Acknowledgements

We thank Dr Willi Maurer for encouraging us to work on this topic and for his many helpful comments throughout the preparation of this manuscript, as well as an associate editor and a referee for their useful suggestions.

References

Abelson, R. and Tukey, J. (1963). Efficient utilisation of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics* **34,** 1347–1369.

Bates, D. and Chambers, J. (1992). Nonlinear models. In *Statistical Models,* S. J. M. Chambers and T. J. Hastie (eds), 421–454. Pacific Grove, California: Wadsworth & Brooks/Cole.

Bates, D. and Watts, D. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Bauer, P. and Hackl, P. (1985). The application of Hunter's inequality in simultaneous testing. *Biometrical Journal* **27,** 25–38.

Branson, M., Pinheiro, J., and Bretz, F. (2003). *Searching for an adequate dose: Combining multiple comparisons and modeling techniques in dose-response studies.* Technical Report No. 2003-08-20, Novartis Pharmaceuticals. Available at `http://www.bioinf.uni-hannover.de/~bretz/paper/TR_MCPMod.pdf`.

Buckland, S., Burham, K., and Augustin, N. (1997). Model selection: An integral part of inference. *Biometrics* **53,** 603–618.

Cox, D. (1977). The role of significance tests. *Scandinavian Journal of Statistics* **4,** 49–70.

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

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. New York: Wiley.

Pinheiro, J. and Bates, D. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing* **6,** 289–296.

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

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

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

Scheffé, H. (1953). A method for judging all contrasts in analysis of variance. *Biometrika* **40,** 87–104.

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Annals of the Institute of Statistical Mathematics* **50,** 1–13.

Tamhane, A., Dunnett, C., and Hochberg, Y. (1996). Multiple test procedures for dose finding. *Biometrics* **52,** 21–37.

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

Westfall, P. and Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* **99,** 25–40.

Received January 2004. *Revised November* 2004.
Accepted December 2004.