Statistical considerations on implementing the MCP-Mod method for binary endpoints in clinical trials

Jingjing Chen*, Tina Liu
Takeda Pharmaceuticals, Cambridge, MA, 02139, USA

A R T I C L E  I N F O
Keywords:
MCP-Mod
Binary endpoint
Dose-response
Dose-finding
Sample size estimation

A B S T R A C T

The Multiple Comparison Procedure – Modelling (MCP-Mod) method was qualified by regulatory agencies (e.g., EMA in 2014 and FDA in 2016) as an efficient statistical method for Phase 2 dose-finding studies when there is uncertainty about dose-response relationship. As this is a relatively new approach, there is limited literature providing practical guidance on its application. In this paper, we evaluated the performance of the MCP-Mod method for clinical trials with a binary primary endpoint, focusing on (1) the impact of sample size, data variability and treatment effect size on the performance of the MCP-Mod, (2) the impact of candidate model misspecification, and (3) optimal sample allocation under a fixed sample size. The evaluation was performed via simulations under different scenarios.

1. Introduction

Developing a new drug and obtaining marketing approval is costly and time-consuming. The average duration of time from clinical testing to product approval is approximately eight years [1]. Approximately half of the Phase 3 trials fail in part because of poor dose selection in Phase 2. Moreover, many new drugs fail to gain approval due to unsatisfactory dosing information rather than failure to demonstrate safety or efficacy [2]. Sacks et al. [2] from the US Food and Drug Administration (FDA) examined 332 marketing applications for therapeutic new molecular entities (NMEs) that were submitted to the Center for Drug Evaluation and Research (CDER) for the first time between October 1, 2000 and September 30, 2012. Their research showed that the most common reason for a first-time drug application failure was the uncertainty in dose selection, defined as uncertainty about the optimal dose to maximize efficacy and to minimize safety risks [2]. Therefore, understanding the dose-response relationship and identifying the optimal dose for Phase 3 confirmatory clinical trials is crucial to the success of the new drug application, and probably the most critical and challenging component of the clinical development program.

The weakness of traditional dose selection methods is well recognized by the biopharmaceutical industry and regulatory agencies. To overcome the shortcomings of traditional dose selection methods, the Multiple Comparison Procedure – Modelling (MCP-Mod) approach, a hybrid approach that combines hypothesis testing and modelling, offers a modelling-based quantitative approach to dose selection with Type I error control. The MCP-Mod method has received wide recognition from the regulatory agencies [3,4]. EMA issued a qualification opinion in 2014 that states the MCP-Mod is an efficient statistical methodology for model-based design and analysis of Phase 2 dose-finding studies. FDA granted the MCP-Mod fit-for-purpose (FFP) designation as an adequate and appropriate method for guiding dose selection for Phase 3 testing in 2016.

The MCP-Mod method was first developed for continuous endpoints, and then was generalized to binary and other types of responses [5]. Some research has been done to evaluate the performance of the MCP-Mod method for continuous and binary endpoints assuming normal distribution (e.g. Ref. [6,7]). However, questions remain regarding the implementation of this method at the trial design stage, given the extent of flexibility and complexity embedded in this method. For example, the MCP-step requires the identification of a set of plausible candidate models and model parameters based on the known pharmacology of the drug and/or that of similar compounds. This is frequently challenging for both statisticians and clinical pharmacologists because of limited knowledge about the compound being studied. These unknowns, if incorrectly specified, may lead to biased inference on the dose selection.

Practical challenges exist to applications of this method to endpoints other than continuous endpoints, given the paucity of information and literatures for them. For instance, the binary endpoint is common in

* Corresponding author.
E-mail address: jingjing.chen@takeda.com (J. Chen).

https://doi.org/10.1016/j.conctc.2020.100641
Received 12 April 2020; Received in revised form 18 July 2020; Accepted 9 August 2020
Available online 14 August 2020
2451-8654/© 2020 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
many therapeutic areas such as gastroenterology and immunology. We were intrigued by these challenges encountered in practice, and evaluated the performance of the MCP-Mod for binary endpoints using simulation studies (under varying assumptions), focusing on (1) the impact of sample sizes, data variability and treatment effects, (2) the impact of candidate model mis-specification, and (3) optimal sample allocation under a fixed sample size. Section 2 reviews the theoretical background of the MCP-Mod method. Sections 3 provides the simulation setup and results evaluating the performance of the MCP-Mod for binary endpoints using 3 metrics: probability of correctly detecting a dose-response signal, probability of correctly identifying the true dose-response model, and probability of correctly identifying minimum effective dose (MED). The paper concludes with a discussion and recommendations on implementation of the MCP-Mod method for binary endpoints in Section 4.

2. Method description

2.1. Traditional dose-ranging study

The main goal of a dose-ranging study is to characterize the dose-response relationship, in order to aid Phase 3 dose-selection. In a traditional dose-ranging study, multiple doses are compared to placebo through pairwise comparisons, and the “best” dose(s) is selected for Phase 3 confirmatory trial. A large number of subjects would be required for each dose arm if the overall Type I error rate were to be properly controlled for pairwise comparisons between active doses and placebo. This may result in an inadequate number of doses in a narrow dose-range in the traditional dose-ranging studies due to feasibility or financial constraints. Furthermore, the pairwise comparisons approach does not make any assumptions about the underlying dose-response relationship. Thus, insufficient number of dose levels and lack of understanding about the dose-response relationship may impact the decision making on selecting a proper dose for Phase 3 clinical trials.

Another category of methodology used in the traditional dose-ranging studies is based on a modelling approach, where a pre-specified parametric model is fitted to the data, e.g., the Emax model. The fitted model is then used to estimate the dose to achieve the desired treatment effect. This approach overcomes the shortcoming of not making any assumptions on the underlying dose-response relationship in the pairwise comparison approach. However, the validity of this modelling-based approach heavily relies on choosing the correct dose-response model upfront, which is often unknown at the design stage.

2.2. MCP-Mod method

Given the shortcomings of the pairwise comparison approach and the modelling-based approach of traditional dose-ranging studies, Bretz, Pinheiro and Branson [7,8,9] proposed the MCP-Mod method which combines the principles of multiple comparisons with modelling techniques. The MCP-Mod method consists of a design stage and an analysis stage.

In the design stage, a set of plausible candidate dose-response models is selected based on the known pharmacology of the drug and/or that of similar compounds. This candidate set is recommended to cover a sufficiently large and diverse set of dose-response shapes to account for model uncertainty. Most parametric dose-response models are listed below:

| Model | \( f(d, \theta) \) | \( f^*(d, \theta^*) \) |
|-------|------------------|------------------|
| Linear | \( E_0 + \theta_0 d \) | \( d \) |
| Linear log-dose | \( E_0 + \theta_1 \log(d + c) \) | \( \log(d + c) \) |
| Exponential | \( E_0 + E_1 \left( \exp\left( \frac{d}{\theta_2} \right) - 1 \right) \) | \( \exp\left( \frac{d}{\theta_2} \right) - 1 \) |
| Emax | \( E_0 + E_{\text{max}} d/ E_{\text{max}} + d \) | \( d/ E_{\text{max}} + d \) |
| Sigmoid EMax | \( E_0 + E_{\text{max}} d^2 / (E_{\text{max}}^2 + d^2) \) | \( d^2 / (E_{\text{max}}^2 + d^2) \) |
| Logistic | \( E_0 + E_{\text{max}} \left( 1 + \exp\left( E_{\text{max}} - d/\theta_2 \right) \right) \) | \( 1 / \left( 1 + \exp\left( E_{\text{max}} - d/\theta_2 \right) \right) \) |
| Quadratic | \( E_0 + \beta_2 d + \beta_3 d^2 \) | \( d + \left( \beta_2 / \beta_3 \right) d^2 \) |
| Beta | \( E_0 + E_{\text{max}} \beta (0, \beta_0 \left( \frac{d}{\beta} \right)^\nu \left( 1 - \frac{d}{\beta} \right)^\delta \) | \( \left( \frac{d}{\beta} \right)^\nu \left( 1 - \frac{d}{\beta} \right)^\delta \) |

The analysis stage includes the MCP-step and the Mod-step.

- The MCP-step is used to assess the presence of a dose response signal (i.e., a non-flat signal) using a trend test from a set of pre-specified plausible candidate models.

\[ T_{\text{max}} = \max_i T_i > q \]

where q is an appropriate multiplicity adjusted critical value, and \( T_i = \frac{cP^{\sum_{i=1}^{K} r_i^2 / n_i}}{\sqrt{\sum_{i=1}^{K} r_i^2 / n_i}} \) is the contrast test for candidate model s. The dose response signal is confirmed if at least one single contrast is statistically significant while the familywise Type I error is controlled. Subsequently, the best dose-response model with a statistically significant dose-response signal is selected based on the maximum contrast test from the pre-specified plausible candidate models via different model selection criteria, e.g., Akaikie Information Criterion (AIC) or Bayesian Information Criterion (BIC) or the most significant contrast test.

- The Mod-step relies on parametric modelling or model averaging to find the ‘optimal’ dose (e.g., minimum effective dose [MED], ED\(_p\)) based on the selected best dose-response model from the MCP-step. MED is defined as the lowest dose ensuring a clinically relevant and statistically significant improvement,

\[ \text{MED} = \min\{ d \in (d_1, d_2) : f(d) > f(d_1) + \Delta \} \]

where \( \Delta \) is the clinical relevance threshold. ED\(_p\) is defined as the lowest dose that gives a certain percentage p of maximum effect \( \delta_{\max} \).

\[ \text{ED}_p = \min\{ d \in (d_1, d_2) : f(d) > f(d_1) + p\delta_{\max} \} \]

For binary outcomes, the dose-response candidate model \( f(d, \theta) \) is defined as a logit link function, \( \text{logit} \{ \text{Prob} (Y = 1) \} = f(d, \theta) \), which allows to interpret the dose-response for binary outcomes through exponential coefficients [5]. Subsequently, the contrast for the trend test in the MCP-step is also defined based on the logit.

The MCP-Mod method is found to be more effective than the pairwise comparisons because of its ability to utilize all available data from the continuum of active doses and placebo to estimate a parametric dose-response curve, which allows for interpolation and extrapolation of effect across a range of doses. Commercial statistical software has been created for implementation of the MCP-Mod method, such as EAST Software by Cytel and ADAPTPLAN by Icon. The DoseFinding package in R is also available to the public.

3. Simulations

The simulation studies were conducted to evaluate the performance of the MCP-Step and the Mod-Step of the MCP-Mod method for binary outcomes in clinical trials. We aimed to evaluate

1. the properties and differential effect of different sample sizes on the performance of MCP-Mod;
3.1. Simulation outline

3.1.1. Binary outcome data simulation

The binary response outcome data was simulated under six common dose-response models as the true underlying dose-response models via 10,000 iterations to mimic 10,000 clinical trials with binary outcomes. The shape of the true underlying dose-response model includes Emax, linear, exponential, logistic, sigmoid Emax, and quadratic models (see Fig. 1). Table 1 shows the parameter specifications of the true underlying dose-response models used in the simulations.

A total of 5 treatment doses \( D \) were considered: placebo (0 mg) and 4 active doses (0.05 mg, 0.2 mg, 0.6 mg, and 1 mg) with 1:1:1:1:1 ratio. The placebo response rate was set at 20%. The maximum treatment effect was set at 60%. A range of sample sizes (i.e., \( N = 10, 20, 40, 80 \) per arm) was considered in the simulation studies.

3.1.2. MCP-Mod analysis

In order to perform the MCP-Mod analysis, a set of pre-specified plausible candidate dose-response models was composed of six common dose-response models, including Emax, linear, exponential, logistic, sigmoid Emax, and quadratic. Two cases were evaluated in the simulation studies:

- Case 1: The true underlying dose-response model was included in the set of plausible candidate models, e.g., six candidate models in the pre-specified candidate set.
- Case 2: the set of pre-specified plausible candidate models was misspecified by omitting the true underlying dose-response, e.g. five candidate models by excluding the true underlying dose-response model in the pre-specified candidate set.

In the MCP-step, the “best” model was identified among a set of candidate models with statistically significant dose-response signal by minimum AIC. The performance of the MCP-step was evaluated via probability of detecting a dose-response signal and probability of correctly identifying the true underlying dose-response model. In particular, probability of correctly detecting a dose-response signal (e.g., non-flat shape or successful proof of concept) was computed as the proportion of detecting a dose-response signal regardless of identifying the correct true underlying dose-response model from 10,000 simulated trials; and probability of correctly identifying the true underlying dose-response model was computed as the proportion of correctly identifying the true underlying dose-response model from 10,000 simulated trials.

In the Mod-step, the MED was estimated as follows and corresponds to \( \text{MED}_2 \) in Bretz et al. [8],

\[
\hat{\text{MED}} = \min \{ d \in \{ d_1, d_6 \} : \hat{f}(d) > f(d_1) + \Delta, L_L(d) > f(d_1) \},
\]

where \( \hat{f}(d) \) is the predicted mean response at dose \( d \), and \( L_L(d) \) is the corresponding lower bound of the pointwise confidence intervals of level 1-2\(\gamma\). The performance of the Mod-step was evaluated via the probability of correctly estimating the MED, where the final MED was estimated based on a weighted estimate across all significant dose-response models selected from the MCP-step via weighted inverse AIC approach on a discrete scale, regardless of identification of true dose-response model. Of note, MED estimation on a discrete scale allows to restrict on the doses within the dosing interval under investigation and avoids problems from extrapolating beyond the dose range under investigation.

3.1.3. Optimal sample allocation

The study team often faces the challenge of a limited study budget.

![Figure 1](image-url)  
**Fig. 1.** Candidate dose response curves.
The study team then has to determine how to optimally allocate the samples with a fixed budget (i.e., a fixed sample size), testing more doses with fewer subjects per arm versus testing fewer doses with more subjects per arm. A separate simulation study was performed to evaluate the impact of sample allocation on the performance of the MCP-Mod method with a fixed total sample size of 120 subjects through 10,000 simulation runs.

- Scenario 1: 5 active doses (0.2 mg, 0.4 mg, 0.6 mg, 0.8 mg, and 1 mg) and placebo with 20 subjects per arm.
- Scenario 2: 3 active doses (0.3 mg, 0.7 mg, and 1 mg) and placebo with 30 subjects per arm.

The set of pre-specified plausible candidate dose-response models included six common dose-response models (Emax, linear, exponential, logistic, sigma Emax, and quadratic) for both Scenario 1 and Scenario 2. Probability of correctly detecting a dose-response signal and probability of correctly identifying the true dose-response model were evaluated.

### 3.2. Simulation results

#### 3.2.1. Metrics for case 1 under various sample size

Table 2 shows the probability of correctly detecting a non-flat dose-response signal, probability of correctly identifying the true underlying dose-response model, and probability of correctly estimating the MED for different sample sizes and different true underlying dose-response models for Case 1. The true underlying model was included in the set of six pre-specified plausible candidate models. It was apparent that allocating more subjects per arm led to higher probability of correctly detecting a dose-response signal, higher probability of correctly identifying the true dose-response model, and higher probability of correctly estimating the MED. Given the simulation setting in this work, the simulation results showed that N = 20 subjects per arm was adequate to detect a non-flat dose-response signal or proof of concept (PoC) for all true underlying dose-response models. However, it may not guarantee correct identification of the true underlying dose-response model. Among all true underlying dose-response models, quadratic and linear models showed a relatively high probability of being correctly selected as the best model from the MCP-step, while sigmoid Emax model was the least likely to be selected even with a large sample size of N = 80 subjects per arm. Depending on the true underlying dose-response model, the sample size needed for correctly estimating the true MED varied. In general, more subjects are needed to precisely select the true MED than the demonstration of detecting a dose-response relationship, especially when the true underlying model is linear or quadratic. It is worth noting that the MED could be still correctly selected when the true underlying dose-response model was not correctly identified as the “best” model from the MCP-step, if the mistakenly identified “best” dose-response model had a similar shape to the true underlying dose-response shape.

Table 3 shows the probability of being selected as the “best” model from the MCP-step under different true underlying dose-response models by different sample sizes in Case 1. It is interesting to note that linear model has the highest probability of being selected as the best model in all scenarios regardless of the true underlying dose-response model. On the contrary, the sigmoid Emax model was least likely to be selected as the best model even if it was the true underlying dose-response model. The likelihood of correctly identifying the true underlying model generally went up as the sample size increased with exception of sigmoid Emax model.

#### 3.2.2. Metrics for case 2 under various sample size

Table 4 shows the probability of correctly detecting a non-flat dose-response signal and the probability of correctly estimating the MED by sample size for different true underlying dose-response models in Case 2, when the set of pre-specified plausible candidate model was misspecified by excluding the true underlying model. For example, when the true underlying dose-response model was Emax model, the selected set of pre-specified candidate model erroneously omitted Emax model by only including linear, Linear, Exponential, Logistic, sigmoid Emax, and quadratic dose-response models. When sample size was relatively small (e.g., N = 10 subjects per arm), the simulation results showed a loss in power in detecting the dose-response signal especially when the true underlying model was quadratic model, given that the shape of quadratic model could be very different from the dose-response models included in the set of plausible candidate models. With a moderate sample size of at least N = 20 subjects per arm, the MCP-step appeared generally having no problems in detecting a non-flat dose-response signal or PoC if a similar dose-response shape to the true underlying dose-response shape was included in the set of candidate models. However, the probability of correctly selecting the MED could be much impacted when the true underlying dose-response model was omitted from the set of candidate models in all scenarios.

#### 3.2.3. Comparison of metrics under case 1 vs. case 2

Fig. 2 presents boxplots comparing the estimated MED from the Mod-step with inclusion of the true underlying dose-response model (Case 1) versus that without inclusion of the true underlying dose-response model (Case 2) in a side-by-side fashion. It was obvious that allocating less subjects per arm led to bigger data variation and could reduce the likelihood of correctly estimating the MED in both Case 1 and Case 2. Certain true underlying dose-response models seemed to require bigger sample size for more precise estimation of the MED such as Emax and Exponential. There appeared to be a power loss with bigger data variation in the estimated MED for all scenarios when the set of candidate models was mis-specified by excluding the true underlying model.

#### 3.2.4. Optimal sample allocation

Table 5 shows the probability of correctly detecting a non-flat dose-response signal by different number of active doses and by true underlying dose-response models with a fixed total sample size.

The simulations results showed little difference in terms of detecting a non-flat dose-response signal between more active doses with fewer subjects per arm.

---

**Table 2**

| True Underlying Model | Probability of Correctly Detecting Dose-Response Signal (%) | Probability of Correctly Identifying True Model (%) | Probability of Correctly Estimating MED (%) |
|-----------------------|----------------------------------------------------------|-------------------------------------------------|--------------------------------------------|
|                       | N = 10 | N = 20 | N = 40 | N = 80 | N = 10 | N = 20 | N = 40 | N = 80 | N = 10 | N = 20 | N = 40 | N = 80 |
| Emax                  | 69.5   | 99.1   | 100    | 100    | 19.0   | 41.6   | 61.3   | 74.1   | 52.6   | 80.3   | 89.8   | 97.1   |
| Linear                | 70.8   | 99.2   | 100    | 100    | 55.7   | 80.6   | 80.7   | 81.4   | 30.8   | 49.6   | 62.4   | 74.5   |
| Exponential           | 67.1   | 98.4   | 99.9   | 100    | 30.9   | 61.7   | 82.4   | 83.0   | 55.7   | 71.3   | 77.7   | 87.2   |
| Logistic              | 87.4   | 99.9   | 100    | 100    | 2.7    | 17.8   | 39.3   | 51.9   | 70.8   | 86.2   | 93.5   | 98.2   |
| Sigmoid               | 79.1   | 99.7   | 100    | 100    | 0.1    | 0.02   | 1.9    | 8.5    | 55.7   | 67.8   | 70.5   | 72.8   |
| Quadratic             | 52.0   | 96.9   | 100    | 100    | 42.3   | 87.9   | 96.1   | 99.5   | 33.2   | 61.2   | 49.8   | 59.0   |

Note: The set of six prespecified plausible candidate model was composed of Emax, Linear, Exponential, Logistic, Sigmoid Emax, and Quadratic. N represents the number of subjects per arm.
correctly included in the candidate set of plausible dose-response ranging studies in practice, we reviewed and evaluated the properties for binary outcomes under different scenarios.

and performance of the MCP-step and Mod-step of the MCP-Mod method facilitate the candidate model specifications at the design stage of dose-

icians to obtain a better understanding of the MCP-Mod method and to assist statisti-

operating characteristics of the MCP-Mod method in order to properly select the best/optimal model regardless of the true underlying dose-

response model when sample size is limited (e.g., N ≥ 10 subjects per arm), as the MCP-step deals with the trade-off between the goodness of fit of the model and the complexity of the model. The accuracy of best/optimal model selection can be improved when sample size increases (e.g., N ≥ 20 subjects per arm).

It is worth noting that sigmoid Emax model is most likely not to be selected among all dose-response models evaluated in the MCP-step of this work, even when sigmoid Emax model is the true underlying model. However, the mis-selection of best/optimal model in the MCP-step does not seem to affect the selection of MED in the Mod-step much. In other words, even if the true dose-response model is missed by the MCP-step, the MED can still be reasonably estimated in the Mod-step much. In other words, even if the true dose-response model is selected among all dose-response models evaluated in the MCP-step, as the MCP-step deals with the trade-off between the goodness of fit of the model and the complexity of the model. The accuracy of best/optimal model selection can be improved when sample size increases (e.g., N ≥ 20 subjects per arm).

The simulation results suggest when the true underlying model is Emax, Linear, Exponential, Logistic, sigomoid Emax, and Quadratic excluding the true underlying dose-response models.

The MCP-step tends to pick the simplest model (e.g., linear model) as the true underlying dose-response model.

The MCP-step does not seem to require a large sample size to detect a non-flat dose-response signal in all scenarios given the simulation setup. However, a larger sample size does improve the accuracy of identifying the correct dose-response model and the precision of estimating the MED for Case 2.

The advances in statistical software technology make it convenient for statisticians to implement the MCP-Mod method in clinical trials. However, there are often many unknowns related to the candidate dose-response models or candidate model specifications in the early stage of drug development. Thus, it is essential to understand the properties and operating characteristics of the MCP-Mod method in order to properly implement it and interpret the results in practice. To assist statisticians to obtain a better understanding of the MCP-Mod method and to facilitate the candidate model specifications at the design stage of dose-ranging studies in practice, we reviewed and evaluated the properties and performance of the MCP-step and Mod-step of the MCP-Mod method for binary outcomes under different scenarios.

The simulation results suggest when the true underlying model is correctly included in the candidate set of plausible dose-response models.

4. Discussion and recommendation

The probability of being selected best model from the MCP-step for different true underlying dose-response models for Case 1.

Table 3

| Sample Size | True Underlying Model | Probability of being Selected Best Model from MCP-Step (%) |
|-------------|------------------------|----------------------------------------------------------|
|             | Emax                   | Linear         | Exponential | Logistic | sigEmax | Quadratic |
| N = 10      | 19.0                   | 33.2           | 0           | 0        | 0.2     | 14.4      |
| Linear      | 3.6                    | 55.7           | 7.1         | 1.2      | 0.05    | 3.2       |
| Exponential | 0.5                    | 35.3           | 30.9        | 0.2      | 0.01    | 0.2       |
| Logistic    | 2.6                    | 69.5           | 1.0         | 2.7      | 0.02    | 10.6      |
| Sigmoid Emax| 4.0                    | 63.6           | 3.2         | 1.8      | 0.05    | 0         |
| Quadratic   | 4.2                    | 4.4            | 0.03        | 0.8      | 0.08    | 42.3      |
| N = 20      | 41.6                   | 32.7           | 0           | 0        | 0.7     | 20.9      |
| Linear      | 4.6                    | 80.6           | 8.8         | 1.1      | 0.08    | 4.0       |
| Exponential | 0.2                    | 36.3           | 61.7        | 0.03     | 0      | 0.1       |
| Logistic    | 1.9                    | 61.4           | 0.7         | 17.8     | 0.05    | 18.1      |
| Sigmoid Emax| 4.8                    | 78.7           | 2.6         | 3.5      | 0.02    | 0         |
| Quadratic   | 7.1                    | 0.8            | 0           | 1.1      | 0.08    | 87.9      |
| N = 40      | 61.3                   | 13.4           | 0           | 0        | 4.1     | 19.0      |
| Linear      | 4.8                    | 80.7           | 7.6         | 1.1      | 0.3     | 5.4       |
| Exponential | 0.01                   | 15.3           | 82.4        | 0.02     | 0      | 2.2       |
| Logistic    | 1.2                    | 34.9           | 0.02        | 39.3     | 6.4     | 18.3      |
| Sigmoid Emax| 5.7                    | 69.9           | 3.2         | 7.1      | 1.9     | 0         |
| Quadratic   | 2.8                    | 0.02           | 0           | 1.1      | 0.03    | 96.1      |
| N = 80      | 74.1                   | 20.8           | 0           | 0        | 10.0    | 12.7      |
| Linear      | 4.9                    | 81.4           | 5.7         | 1.2      | 0.5     | 63.3      |
| Exponential | 0                      | 2.5            | 83.0        | 0.1      | 0      | 14.4      |
| Logistic    | 0.4                    | 11.3           | 0           | 51.9     | 24.0    | 12.4      |
| Sigmoid Emax| 6.4                    | 58.7           | 0.3         | 7.8      | 8.5     | 0         |
| Quadratic   | 0.3                    | 0              | 0           | 0.2      | 0      | 99.5      |

Note: The set of five candidate model was composed of Emax, Linear, Exponential, Logistic, sigmoid Emax, and Quadratic excluding the true underlying dose-response model.

N represents the number of subjects per arm.

Table 4

| True Underlying Model | Probability of Detecting Dose-Response Signal (%) | Probability of Correctly Estimating MED (%) |
|-----------------------|---------------------------------------------------|-------------------------------------------|
| N = 10                | N = 20                | N = 40              | N = 80              | N = 10                | N = 20                | N = 40              | N = 80              |
| Emax                  | 64.0                  | 98.2                | 100                 | 100                  | 46.3                  | 79.5                | 89.7                | 97.0                |
| Linear                | 68.4                  | 98.9                | 100                 | 100                  | 28.5                  | 46.7                | 61.2                | 75.5                |
| Exponential           | 52.8                  | 97.0                | 99.9                | 100                  | 35.8                  | 50.2                | 54.4                | 60.2                |
| Logistic              | 85.4                  | 99.9                | 100                 | 100                  | 57.2                  | 71.4                | 82.2                | 92.5                |
| Sigmoid Emax          | 78.0                  | 99.8                | 100                 | 100                  | 50.6                  | 62.9                | 63.1                | 62.7                |
| Quadratic             | 18.3                  | 49.9                | 92.8                | 99.9                | 12.8                  | 17.5                | 18.4                | 19.9                |

Sample size per arm versus less active doses with bigger sample size per arm. However, sample allocation per arm appeared to have an impact on the precision of the MED estimation. Allocating more samples per arm generally led to better estimation of the MED with the exception of Emax as the true underlying dose-response model.
When the candidate set of dose-response models is mis-specified by neglecting the true underlying dose-response model, the probability of detecting a non-flat dose-response shape or establishing PoC does not seem to be affected much in the MCP-step, if a similar dose-response shape as the true shape of dose-response model is included in the candidate set. However, the precision of estimating the MED in the Mod-step may be comprised. Given that some dose-response models are highly correlated, a general rule of thumb in practice is to cover a variety of dose-response shapes with the least number of candidate models, so that the model mis-specification issue can be adjusted by the MCP-step and Mod-step to make the results not deviate from the truth too much. Additionally, the MED can be estimated from an identified single “best” dose-response model or alternatively estimated as a weighted estimate of MED across all significant dose-response models via model averaging techniques such as weighted inverse AIC approach. The MED can be estimated on a continuous scale or discrete scale.

The simulation results also show that the MCP-Mod method is more related to the total sample size. When the budget or the total sample size is fixed, the number of active dose levels doesn’t have a great impact on the performance of detecting a non-flat dose-response signal or establishing PoC, as compared to allocating more subjects to each dose with less active doses. However, allocating more subjects to each dose does improve the precision of estimating MED. It is worth noting that the sample size estimated based on the MCP-Mod method in existing software is targeted to detect the existence of dose-response relationship. In order to have a high probability of correctly identifying the MED to move forward to the Phase 3 trial, the sample size needs to be larger. We would recommend using the modelling and simulation techniques for sample size and power estimation for objectives other than PoC. We recommend taking the study objectives into consideration when determining the sample size of a Phase 2 dose-finding study.

Our simulations are limited to the scenarios and cases studied in this work. We only considered the case of binary outcome data in this evaluation, although a similar conclusion would be expected for continuous and time-to-event data. In addition, we only presented one effect size (placebo rate = 20%, maximum treatment effect = 60%) in the article, given that different effect sizes show similar trend. Caution should be taken when applying the MCP-Mod method in small sized trials, where the power of correctly identifying best/optimal model or precisely estimating the MED could be compromised. Further research on the MCP-Mod method applying to co-primary binary endpoints in certain therapeutic areas would be helpful.
Acknowledgement

This work was sponsored by Takeda. We thank Lu Li (Johns Hopkins University) for her assistance in this work and Dr Xiaopan Yao for the useful discussions in the summer of 2018. We also thank Dr. Sheela Kolluri for providing valuable comments in writing this manuscript.

References

[1] K.I. Kaitin, J.A. DiMasi, Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000-2009, Clin. Pharmacol. Ther. 89 (2) (2011) 183–188.
[2] L. V. Sacks, H. H. Shamsuddin, Y. I. Yasinskaya, K. Bouri, M. L. Lanthier, R. E. Sherman, Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000–2012, J. Am. Med. Assoc. 311 (4) (2014) 378–384.
[3] European Medicines Agency (EMA), Qualification Opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of phase II dose finding studies under model-uncertainty, 23 January, https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-mcp-mod-efficient-statistical-methodology-model-based-design-analysis-phase-ii_en.pdf, 2014.
[4] USA Food and Drug Administration (FDA), Statistical review and evaluation qualification of statistical approach: MCP-Mod, in: https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/UCM508701.pdf, 2016.
[5] J. Pinheiro, B. Bjørnmark, E. Glimm, F. Bretz, Model-based dose finding under model uncertainty using general parametric models, Stat. Med. 33 (2013) 1646–1661.
[6] S. Wang, Discussion of the “White Paper of the PhRMA working group on adaptive dose-ranging designs”, J. Biopharm. Stat. 17 (2007) 1015–1020.
[7] Y. Zhou, S. Chen, D. Sullivan, Y. Li, Y. Zhang, W. Xie, H. Zhang, Y. Tang, L. Wang, A. Hartford, B. Yang, Dose-ranging design and analysis methods to identify the minimum effective dose (MED), Contemp. Clin. Trials 63 (2017) 59–66.
[8] F. Bretz, J. Pinheiro, M. Branson, Combining multiple comparisons and modelling techniques in dose-response studies, Biometrics 61 (2005) 738–748.
[9] J. Pinheiro, B. Bjørnmark, F. Bretz, Design and analysis of dose-finding studies combining multiple comparisons and modelling procedures, J. Biopharm. Stat. 16 (5) (2006) 639–656.