A RESAMPLING BASED APPROACH IN EVALUATION OF DOSE-RESPONSE MODELS

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Min Fu
December, 2014

Examining Committee Members:

Richard M. Heiberger, Advisory Chair, Statistics
Sanat K. Sarkar, Statistics
Xu Han, Statistics
José C. Pinheiro, Quantitative Science, Janssen R&D US
Bo Ji, External Reader, Computer and Information Science

ABSTRACT

A RESAMPLING BASED APPROACH IN EVALUATION OF
DOSE-RESPONSE MODELS

Min Fu
DOCTOR OF PHILOSOPHY

Temple University, December, 2014

Professor Richard M. Heiberger, Chair

In this dissertation, we propose a computational approach using a resampling based permutation test as an alternative to MCP–Mod (a hybrid framework integrating the multiple comparison procedure and the modeling technique) and gMCP–Mod (generalized MCP–Mod) [11], [29] in the step of identifying significant dose-response signals via model selection. We name our proposed approach RMCP–Mod or gRMCP–Mod correspondingly. The RMCP–Mod/gRMCP–Mod transforms the drug dose comparisons into a dose-response model selection issue via multiple hypotheses testing, an area where not much extended researches have been done, and solve it using resampling based multiple testing procedures [38]. The proposed approach avoids the inclusion of the prior dose-response knowledge known as "guesstimates" used in the model selection step of the MCP–Mod/gMCP–Mod framework, and therefore reduces the uncertainty in the significant model identification.

When a new drug is being developed to treat patients with a specified disease, one of the key steps is to discover an optimal drug dose or doses that would produce the desired clinical effect with an acceptable level of toxicity. In order to find such a dose or doses (different doses may be able to produce the same or better clinical effect with similar acceptable toxicity), the underlying dose-response signals need to be identified and thoroughly examined through statistical analyses. A dose-response signal refers to the fact that a drug has different clinical effects at many quantitative dose levels. Statistically speaking, the dose-response signal is a numeric relationship curve (shape) between drug doses and the clinical effects in quantitative measures. It's often been a challenge to find correct and accurate efficacy and/or safety dose-response signals that would best describe the dose-effect relationship in the drug development process via conventional statistical methods because the conventional methods tend to either focus on a fixed, small number of quantitative dosages or evaluate multiple pre-defined dose-response models without Type I error control. In searching for more efficient methods, a framework of combining both multiple comparisons procedure (MCP) and model-based (Mod) techniques acronymed MCP-Mod was developed by F. Bretz, J. C. Pinheiro, and M. Branson [11] to handle normally distributed, homoscedastic dose response observations. Subsequently, a generalized version of the MCP–Mod named gMCP–Mod which can additionally deal with binary, counts, or time-to-event dose-response data as well as repeated measurements over time was developed by J. C.Pinheiro, B. Bornkamp, E. Glimm and F. Bretz [29]. The MCP–Mod/gMCP–Mod uses the "guesstimates" in the MCP step to pre-specify parameters of the candidate models; however, in situations where the prior knowledge of the dose-response information is difficult to obtain, the uncertainties could be introduced into the model selection process, impacting on the correctness of the model identification.

Throughout the evaluation of its application to the hypothetical and real study examples as well as simulation comparisons to the MCP–Mod/gMCP–Mod, our proposed approach, RMCP–Mod/gRMCP–Mod seems a viable method that can be used in the practice with some further improvements and researches that are still needed in applications to broader dose-response data types.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Heiberger for his guidance and his willingness to work with me on a variety of computationally statistical issues and other related technical challenges that I have encountered through out this dissertation research. Very sincere thanks to Dr. Sarkar for his rigorous advice on the research basics and Dr. Han for his help through sensible questions. An appreciation goes to Dr. José Pinheiro, a mentor of my professional life as a statistician, without him I would not be where I am now. Sincere thanks to all the committee members including the external reader for spending hours reading through this document.

This research was supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854.

Special thanks goes to my colleague Chyi-Hung Hsu. Without his expertise in helping using R on the high performance computer cluster, I would not have had finished all the computations with big struggles.

To my wife, Ying Peng, and daughter, Georgie,
for their unconditional support, love, and faith during
this long journey.

To my mother, Weilian Fei, who taught me the
tenaciousness for the life long learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTON

When a new drug is being developed to treat patients with a specified disease, one of several critical decision-making processes is to determine whether the investigational drug has any dose–response effects (relationship signals). A dose–response effect refers to a quantitative association between a drug dosage and a desired clinical outcome. For example, a meta-analysis performed on a class of cholesterol lowering drugs generically named “statin” showed that the doses ranging from 5 mg to 80 mg daily could monotonically reduce 10% to 60% of the elevated lower-density lipoprotein (LDL) cholesterol concentration in 2 to 6 weeks [25]. On the other hand, there are also drugs that are hardly dose-dependent (no dose–response relationship), such as the angiotensin-converting-enzyme (ACE) inhibitors used for treating patients with hypertension.

The conventional statistical methods used in such a dose–response relationship research have been generally classified into two categories: 1) multiple comparisons approach and 2) modeling approach. In multiple comparisons approach, the drug doses are considered a qualitative factor. The primary goal of the multiple comparisons approach is to identify at least one drug dosage that is significantly better than a control (a placebo or an active). In the dose–response research, a placebo or an active control is usually defined as a 0 dosage in quantification. The placebo often takes a form of drug pills with no drug ingredients included, such as the sugar pills. They are randomly administered to a group of experimental units (patients) as a control group in the blind fashion to protect the randomization schema. It is unconceivable that any biological drug effect would be expected from a placebo, although psychological placebo effect in clinical studies is a common phenomenon. An active (control) refers to a drug that has been successfully developed with some established clinical effects. In order to establish a new treatment, a novel medicine needs to be studied against the placebo and an active control, if there is a similar drug on the market, for better clinical effects or for a safer profile (less toxicity). The limitation of any multiple comparisons procedures used in the dose–response research is the restriction to a small number of discrete dosages. This is partly because of the financial burden on investigating a wide range of doses and partly because of the unknown safety profile of a new investigational medicine. The unknown safety profile often prevents dosing studies from going for riskier dose levels. The modeling approach, on the other hand, takes the drug dosages in a continuous quantitative scale by assuming some functional relationships (shapes) between the dose and the response (clinical effect). The objective of the modeling approach is to discover any quantitative associations (models) that best represent the underlying dose–response relationship. Once a dose–response model is identified, further evaluations of the optimal dosages can be performed using the model. Because the covariates of interest can be included in modeling dose–response effects, the model based approach has a potential impact on personalized treatment strategies in which a drug may be tailored at the variable dose levels for different patients who have different biological response to the drug. The drawback for the model based approach is that without thorough and proper comparisons for many different models with the type I error rate being controlled, the validity of the findings is often highly dependent on the correctness of the assumed model(s).

F. Bretz, J. C. Pinheiro, and M. Branson [11] proposed an approach acronymed MCP–Mod that combines the multiple comparisons procedure (MCP) and the modeling (Mod) technique into a hybrid framework named MCP–Mod. The MCP–Mod takes advantages while addressing disadvantages of both methods. The MCP-Mod is a parametric based method for the dose–response data with the normality assumption. J. C.Pinheiro, B. Bornkamp, E. Glimm and F. Bretz [29] further extended the MCP–Mod to cover a broad range of the data types such as binary, counts, or time-to-event as well as repeated dose–response measurements over time. The extension is referred as generalized MCP–Mod (gMCP–Mod). Both of the MCP–Mod and gMCP–Mod start with specifying a set of selected candidate models (candidate set) and then identify the most significant model(s) that best describe(s) the underlying dose–response relationship reflected in the study observations via an optimal contrasts test which controls the Type I error rate in the multiple comparisons perspective. Once the best model(s) are identified, the target dosage(s) can be estimated from the model(s) for the further drug development.

1.1 Motivation for the Dissertation

As part of the MCP step in MCP–Mod or gMCP–Mod, the optimal contrast(s) over the candidate set of models need to be pre-calculated for the multiple testing to identify significant dose–response model(s) in the candidate set. This calculation of the optimal contrast(s) requires model parameters associated with dose–response shapes in the candidate set to be specified using prior dose–response information or knowledge. In the practice of new drug development process, the prior information for the model parameterization tend to come from previous studies or biologically educated guesses known as the “guesstimates”. These “guesstimates” can potentially contribute to model uncertainties and impact on the control of Type I error rate for which the multiple comparisons of many models are involved. The motivation to find alternatives that could avoid using the “guesstimates” in identifying significant model or models that best describe the dose–response relationships led to the research that is detailed in this dissertation.

The research described in the following sections explores a new, computational approach based on a resampling permutation test to compare the significance of the multiple candidate models. We name it RMCP–Mod/gRMCP–Mod corresponding to MCP–Mog/gMCP–Mod. The proposed approach uses the permutation distribution of the minimum p-values obtained from a permutation test statistic for testing the null hypotheses. These null hypotheses assume no dose–response effect (different doses have no additive effects) between each model compared and a constant (flat) model in the candidate set. Similar to MCP–Mod or gMCP–Mod, RMCP–Mod/gRMCP–Mod starts with defining a candidate set of models. Instead of completely specifying model parameters using any "guesstimates", we fit these models directly to the original study observations as well as to each of the permutation samples that is drawn from the observations without replacement. The permutation process reflects the sense of the null hypothesis of no dose–response effect, i.e., the original dosing levels (labels) to which each experimental unit (patient) randomly assigned are completely exchangeable among all patients. For the original study observations and each permutation sample, we choose a statistic to test a set of null hypotheses that assume each model is no better than a no-dose-response model. The no-dose-response model is basically specified as a flat model that is technically a simple linear model with the intercept term only. We construct a permutation distribution of the test statistic for each fitted model by congregating all permutation test statistic values from all permutation samples. We then obtain a set of observed p-values by comparing their observed test statistics to the corresponding permutation distributions for all candidate models. The observed p-value is actually the quantile of the observed test statistic value from the model fitted to the original study sample on the distribution of the permutation statistic. For each permutation sample, using the same reference distribution, we also calculate a set of permutation p-values. We order these permutation p-values and identify the smallest p-value which represents the best model for each permutation sample. The permutation distribution of the minimum p-values can then be built based on these smallest p-values from all permutation samples. We finally identify the best model(s) by comparing the observed p-values to the pre-specified critical region on the distribution of the minimum p-values, in which the FWER is controlled.

The RMCP–Mod/gRMCP–Mod is a non-parametric statistical method through the computationally intensive technique due to the need of a large numbers of permutation samples drawn from an original study observations. The Computational statistical approaches are increasingly appealing to researchers in the drug development area due to several reasons of:

- less assumptions needed;
- proved resampling-based multiple comparison procedures in controlling Type I error rate or FWER;
- valid alternatives to conducting clinical trials for cost saving;
- warranted feasibility due to availability of modern computational technologies.

1.2 Structure of the Dissertation

This dissertation is organized as follows: an overview of the dose–response analysis techniques is provided via literature reviews in Chapter 2 with an emphasis on the evaluation of basic and theoretic concept of the multiple comparisons procedures, the model-based approaches, and the MCP–Mod or gMCP–Mod frameworks. In Chapter 3, the concept of resampling methods and its applications is introduced. Although the contents could be included as part of literature review, it is felt that the resampling idea plays the fundamental role in developing our proposed approach and a dedicated chapter would provide sufficient descriptions leading into the new method discussion in the follow Chapter. In Chapter 4 we proposed our approach (RMCP–Mod) using a resampling based permutation test as an alternative to the step in MCP–Mod for selecting and evaluating dose–response models. Examples of the new approach applied to the hypothetical and real studies are given. The chapter also reports the results from simulations in statistical power comparisons between the RMCP–Mod and the MCP–Mod on the normal data. Chapter 5 extends such power comparisons to the gRMCP–Mod with gMCP–Mod for the dose–response measures in binary and counts data types. Chapter 6 concludes this dissertation with a summary of the proposed method and with potential extended topics for future research.

CHAPTER 2

LITERATURE REVIEW

In general, decisions through dose–response studies can be divided into two main components: 1) determine whether a new drug has some overall dose–response effect on the clinical outcomes of interest, i.e., dose signal, which is referred to Proof-of-Concept (PoC); 2) estimate (select) a target dose such as the minimum effective dose (MED) defined as the smallest dose with a relevant clinical effect exceeding that of placebo or a control by a threshold that is statistically significant and clinically meaningful. To accomplish 1) and 2), studies known as Phase II clinical trials are often conducted in the drug development process. If the Phase II studies are positive, a couple of similarly designed, Phase III clinical studies are carried out to confirm the drug effects as well as the MED with larger samples sizes. In the following sections we will focus on reviewing commonly used statistical methods in analyzing data collected from phase II clinical trials for the PoC establishment and the MED identification.

2.1 Multiple Comparisons Procedures (MCP)

Conventionally, the overall dose–response signal (PoC) is often analyzed by the statistical trend tests. After the dose–response signal is established, the estimation of a target dose (MED) can be accomplished using the step-wise multiple testing procedures.

2.1.1 Dose–Response Trend Tests

In a simplified description, we assume that the dose–response has the following one-way layout:

$$y_{ij} = \mu_i + \varepsilon_{ij} \tag{2.1}$$

where y_{ij} denotes the observation from the jth subject at the dose level $i; j = 1, \ldots, n_i$, $i = 0, \ldots, k$ with $k \geq 2$; μ_i's denote the mean dose effects, and ε_{ij}'s denote the independent normally distributed errors with mean 0 and variance of σ^2. The index $i = 0$ refers to a zero dose which usually represents the placebo or a control in analyzing clinical trial data. The group mean in each dose group and overall mean are given respectively as below:

$$\bar{y_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},\ \bar{y} = \frac{1}{N}\sum_{i=0}^{k} y_i.$$

and the pooled variance is estimated by

$$s^2 = \frac{1}{v}\sum_{i=0}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y_i})^2.$$

with $v = N-(k+1)$ degrees of freedom, where $N = \sum_{i=0}^{k} n_i$ is the total number of patients.

Investigating any overall dose–response effects can be achieved by testing the null hypothesis of no dose–response effect among these k dose groups:

$$H_0 : \mu_0 = \ldots = \mu_k \tag{2.2}$$

against the restrictive alternative:

$$H_a : \mu_0 \leq \ldots \leq \mu_k \text{ with } \mu_0 < \mu_k, \tag{2.3}$$

note that (2.3) assumes monotonically ordered treatment μ_i (restrictive). If the monotonicity condition holds, powerful tests can be derived.

Single Contrast Test

The contrast tests were first introduced by Abelson and Tukey [1] as well as Schaafsna and Smid [30] in the context of the dose–response testing. These are powerful tests to detect any overall dose–response signals and can be applied to a variety of statistical models, including general linear models allowing for covariates with factorial treatment structures [22].

The test statistic for a single contrast test is given by

$$t = \frac{\sum_{i=0}^{k} c_i \bar{y}_i}{s\sqrt{\sum_{i=0}^{k} \frac{c_i^2}{n_i}}},$$

where $c_0, \ldots, c_k$ denote the fixed constants subject to $\sum_{i=0}^{k} = 0$. Here $c_0, \ldots, c_k$ are known as the contrast coefficients. The test statistic t follows a central t-distribution with v degree of freedom under the null hypothesis H_0 (2.2). When H_0 is not true, t follows a non-central t-distribution with a non-centrality parameter:

$$\tau = \frac{\sum_{i=0}^{k} c_i \mu_i}{\sigma\sqrt{\sum_{i=0}^{k} \frac{c_i^2}{n_i}}}.$$

Numerous proposals for the choice of the contrast coefficients have been made such as those in the article by Tamhane at al [34]. The optimal choice of the contrasts depends on the unknown true mean vector $\mu_0, \ldots, \mu_m$, and particularly on the correlation between the c_i's and the μ_i's, i.e., how the contrast coefficients of c_i closely reflect the patten of the true means of μ_i.

Multiple Contrasts Test

In addition to the single contrast tests described above, the multiple contrasts tests have been used by identifying a set of contrast vectors including misspecified contrast coefficients in the alternative region. The goal of the any multiple contrasts test is to achieve more robust testing results.

Let $q \geq 2$ be the number of contrast vectors, i.e. $\boldsymbol{c_j} = (c_{j0}, \ldots, c_{jk})', j = 1, \ldots, q$, and $t_1, \ldots, t_q$ be the corresponding test statistics, then the multiple contrast test statistic is the maximum statistic value among the t_j's:

$$t_{\max} = \max\{t_1, \ldots, t_q\}. \tag{2.4}$$

For the normal model (2.1), it can be shown that under the null hypothesis (2.2), the $t_1, \ldots, t_q$ are jointly multivariate t-distributed with v degrees of freedom and correlation matrix $\boldsymbol{R} = (\rho_{ij})$, where

$$\rho_{ij} = \frac{\sum_{\ell=0}^{k} \frac{c_{i\ell} c_{j\ell}}{n_\ell}}{\sqrt{\left(\sum_{\ell=0}^{k} \frac{c_{i\ell}^2}{n_\ell}\right)\left(\sum_{\ell=0}^{k} \frac{c_{j\ell}^2}{n_\ell}\right)}}, \quad i, j \leq q, \ \ell = 0, \ldots, k. \tag{2.5}$$

Let $t_\alpha(q, v, \boldsymbol{R})$ denote the upper α equicoordinate critical point of the distribution of the test statistic (2.4), then the multiple contrasts test rejects H_0 if $t_{max} > t_\alpha(q, v, \boldsymbol{R})$.

Bretz et al. [10] provided a list of some multiple comparisons procedures (not necessarily trend tests) which can be formulated as the multiple contrasts tests including many-to-one comparisons, all-pair comparisons, and multiple comparisons with the best.

Extended Multiple Contrasts Test The multiple contrasts tests have been extended to the general linear model with covariates in addition to the dose factor being included in:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.6}$$

where $\boldsymbol{y}$ is an $N \times 1$ observation vector, $\boldsymbol{X}$ is a fixed and known $N \times p$ design matrix, $\boldsymbol{\beta}$ is a fixed and unknown $p \times 1$ parameter vector and $\boldsymbol{\varepsilon}$ is a $N \times 1$ random error vector. The elements (ε_i) of $\boldsymbol{\varepsilon}$ are assumed to be independent and normally distributed random errors with mean of 0 and variance of σ^2. Note that the one-way design of (2.1) is a special case of (2.6) and the dimension of $p = k + 1$ with k does not include zero (zero dose group) in notation here. The usual least square estimates for β_i's and σ^2 are given as:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-}\boldsymbol{X'y} \text{ and } s^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{v}$$

where $v = N-$ rank$(\boldsymbol{X})$ and $(\boldsymbol{X'X})^{-}$ denotes a generalized inverse of $(\boldsymbol{X'X})$. With these setups, we have the pivotal quantity:

$$t_j = \frac{\boldsymbol{c_j'}\hat{\boldsymbol{\beta}}}{s\sqrt{\boldsymbol{c_j'}(\boldsymbol{X'X})^{-}\boldsymbol{c_j}}}, \quad j = 1, \ldots, q, \tag{2.7}$$

where q is the number of single contrast tests. The elements of the $p \times 1$ vectors $\boldsymbol{c_j}$ incorporates the weights of other elements of $\boldsymbol{\beta}$, such as covariates, in addition to the treatment levels as in the case of single contrast tests. The joint distribution of $t_1, ..., t_q$ is also a multivariate t with v degrees of freedom and correlation matrix $\boldsymbol{R} = \boldsymbol{DC'}(\boldsymbol{X'X})^{-}\boldsymbol{CD}$, here $\boldsymbol{C}_{p\times q} = (\boldsymbol{c_1}, \ldots, \boldsymbol{c_q})$ and $\boldsymbol{D} = \text{diag}(\boldsymbol{c_j})'(\boldsymbol{X'X})^{-}\boldsymbol{c_j})^{-1/2}$. If $v \to \infty$ or σ is known, the corresponding multivariate normal distribution holds in the limit.

2.1.2 Estimation of Minimum Effective Dose (MED)

If an overall dose–response signal is established through a contrast test, then the interest turns to estimate or select the minimum effective dose, MED, which is the lowest dose whose effect exceeds that of the placebo or a control by a specified threshold. Consider a primary efficacy endpoint in a dose–response study and assume that higher values of the endpoint indicate an improvement in the medical condition of interest. The dose difference in efficacy measured by group means between investigational dosages and a control can be defined as 1) the mean difference, $\delta_i = \mu_i - \mu_0$, $1 \leq i \leq k$, where dose i is considered effective if the δ_i is greater than a specified threshold value $\delta > 0$; or 2) the ratio of the means, $\lambda_i = \mu_i/\mu_0$ (assuming $\mu_0 > 0$), $1 \leq i \leq k$, where dose i is regarded effective if the ratio, λ_i, exceeds a specified threshold value $\lambda > 1$. The MED can be expressed correspondingly to definition 1):

$$\text{MED} = \min\{i : \mu_i > \mu_0 + \delta\}, \tag{2.8}$$

and to 2):

$$\text{MED} = \min\{i : \lambda_i > \lambda\}. \tag{2.9}$$

We use the MED in the ratio expression to further illustrate the relevant multiple comparisons procedures. A stronger definition of (2.9) can be expressed as

$$\text{MED} = \min\{i : \lambda_j > \lambda\} \text{ for all } i \leq j \leq k. \tag{2.10}$$

If the dose–response signal is monotonic then the definitions of (2.9) and (2.10) are basically the same but (2.9), i.e., without the monotonic assumption, is more of reflecting the practice. The probability of any wrongly declared effective MED needs to be controlled at a specified level α:

$$P(\widehat{\text{MED}} < \text{MED}) \leq \alpha,$$

where $\widehat{\text{MED}}$ is an estimation of the MED. This can be achieved by formulating it into a multiple comparisons problem that can be strongly controlled for FWER:

$$H_{0i} : \mu_i \leq \lambda\mu_0 \text{ vs. } H_{ai} : \mu_i > \lambda\mu_0, \quad 1 \leq i \leq k.$$

If definition (2.9) is used, the $\widehat{\text{MED}}$ would be:

$$\widehat{\text{MED}} = \min\{i : H_{0i} \text{ is rejected}\}.$$

Similarly, when the definition (2.10) is assumed, the $\widehat{\text{MED}}$ would be:

$$\widehat{\text{MED}} = \min\{i : H_{0j} \text{ is rejected for all } i \leq j \leq k\}.$$

The step-down testing procedures based on pairwise contrasts proposed by Tamhane, Hochberg, and Dunnett [34] are popular and practically used in the dose–response researches due to the reasons that:

- when the dose–response is not monotone, only pairwise contrasts yields procedures that still control the FWER (Bauer) [5];

- the procedures are simple to use;
- the procedures can be easily extended to the non-normal data with appropriate two-sample statistics.

For a closed family of hypotheses $\{H_{0i}, 1 \leq i \leq k\}$, it is easy to construct the closed testing procedures. All that is needed are separated α-level tests for the individual H_{0i}, which must be applied in a step-down manner to control the FWER. A hypothesis H_{0i} is tested and rejected at level α if and only if all the hypotheses H_{0j} are significant at level α for $j \geq i$. The following pairwise t-statistic

$$t_i = \frac{\bar{y}_i - \lambda \bar{y}_0}{s\sqrt{\lambda^2/n_0 + 1/n_i}} \tag{2.11}$$

is used for testing $H_{01}, \ldots, H_{0k}$.

If a monotonicity can be assumed in the dose–response effect, then a simplified closed step-down procedure can be used to test hypotheses, $H_{0k}, \ldots, H_{01}$, in a stepwise manner, each at level α. The testing procedure rejects H_i if all the hypotheses H_j for $j > i$ are rejected at $t_i > t_\alpha(v)$, where $t_\alpha(v)$ is the upper α critical point of the t-distribution with v degrees of freedom. Otherwise, the test procedure stops and retain all the hypotheses H_j for $j \leq i$. Refer to Tamhane et al. (1996) [34] and Dmitrienko et al. (2010) [14] for detailed discussions and examples.

Without the monotonicity assumption, the ordinary step-down procedure uses the joint distribution of the corresponding union-intersection statistic for testing $H_i' = \bigcap_{j=1}^{i} H_j$:

$$t_{i,\max} = \max_{1<j\leq i} t_j,$$

where t_j is the pairwise t test statistic (2.11). The join distribution of t_js is a multivariate t-distribution with $v = N - (k+1)$ degrees of freedom and correlation matrix $\boldsymbol{R}_k = \{\rho_{ij}\}$ with a product correlation structure, i.e., $\rho_{ij} = \tau_i \tau_j$, where

$$\tau_i = \frac{\lambda}{\sqrt{\lambda^2 + r_i}}, \quad r_i = \frac{n_0}{n_i}, \quad 1 \leq i \leq k.$$

In a balanced case, when $n_1 = \ldots = n_k = n$ and $r = n_0/n$, the off-diagonal elements in the correlation matrix $\boldsymbol{R}$ are given by $\rho_{ij} = \lambda^2/(\lambda^2 + r), i \neq j$. Let $t_\alpha(k, v, \boldsymbol{R}_k)$ denote the upper α equicoordinate critical point of the k variate t-distribution under the null hypotheses with correlation matrix $\boldsymbol{R}_k$. These critical points can be computed using the algorithms in Genz and Bretz (2009) [18]. The procedure rejects H'_j if and only if $H'_k, \ldots, H'_{j+1}$ are rejected.

In above step-down procedures (with or without monotonicity assumption), the dose corresponding to the last rejected hypothesis is the identified MED.

2.2 Modeling Approaches (Mod)

The modeling approach is another commonly used method for analyzing dose–response data. It pre-specifies a parametric model which assumes a functional relationship between the dose in a numerically continuous scale and the response. The model is fitted to the study observations and evaluated through the goodness of fit measures. The fitted model can then be used to estimate the MED, if the model is believed to be the one that closely expresses the underlying dose–response relationship through the evaluation. Such an approach provides the flexibility in investigating (interpolating) the dose effects beyond the limited dose levels being experimented in patients. For example, in an experiment, for a variety of reasons, the drug doses afforded to be studied are 20mg, 40mg, and 100 mg. However, the MED may truly lie between 40mg and 100mg. The modeling approach would have the potential to closely estimate the MED whereas the multiple comparison procedures could only arrive at a decision of choosing either 40mg or 100mg.

The general framework for the modeling statistical analysis is that the dose–response $\boldsymbol{Y}$(which can be an efficacy or a safety outcome) is observed from a given set of grouped subjects in parallel. Each of these subjects is randomly assigned to one of the dosage groups of $d_1, d_2, ..., d_k$ plus a control d_0 for a total of $k+1$ treatment arms. For the purpose of testing the PoC and estimating the MED, the one way layout similar to (2.1) is usually used for the model specification:

$$y_{ij} = \mu_{d_i} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \text{iid } N(0, \sigma^2),\ i = 0, 1, \ldots, k, j = 1, 2, \ldots, n_i, \quad (2.12)$$

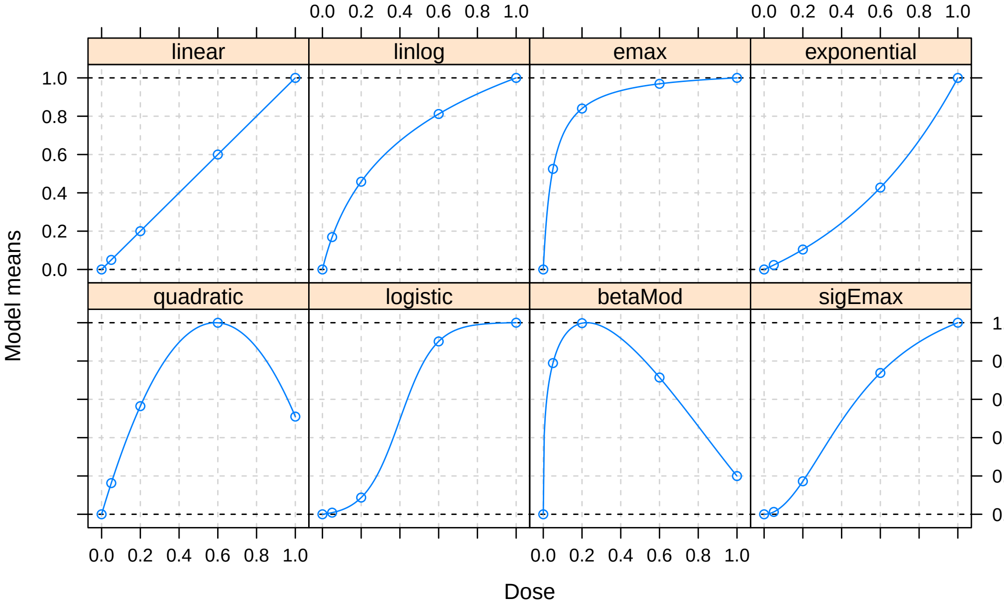
where n_i is the sample size of the dose group i; the mean response at dose d_i can be represented as $\mu_{d_i} = f(d_i, \boldsymbol{\theta^0})$ for some dose–response model $f(\cdot)$ parameterized by a vector of parameters $\boldsymbol{\theta^0}$; and ε_{ij} is the error term for the patient j in the dose group i.

To estimate the model parameter θs, the ordinary least squares (OLS) estimates that minimize the residual sum of squares $\sum_{i=1}^{k} \sum_{j=1}^{n_i} |y_{ij} - f(d_i, \boldsymbol{\theta^0})|^2$ are typically applied, under the assumption of the independently and identically distributed error terms ε_{ij}. In the case of any nonlinear dose–response models, the nonlinear least squares algorithms are needed for estimating $\boldsymbol{\theta^0}$. The most popular one is the Gauss-Newton algorithm (Bates and Watts, 1988; Seber and Wild, 1989)[4],[32]. The algorithm is an iterative procedure to solve (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, so-called initial values, for the parameters. Methods for deriving the initial estimates for the nonlinear models are also discussed in Bates and Watts (1988)[4]. The Gauss-Newton algorithm for the nonlinear least squares can be found in the mainstream of statistical software packages. It is implemented in the functions *nls* (Chambers and Hastie, 1992) [13] and *gnls* (Pinheiro and Bates, 2000) [28] of S-PLUS and R as well as in the SAS procedure, PROC NLIN (Freund and Littell, 2000) [17]. Examples on the use of these functions and procedures can be found in these references.

The commonly used dose–response models in drug research area are *linear*, *linear-in-log-dose*, *quadratic*, *exponential* (power), *beta*, E_{max}, and *logistic* models, etc.. Detailed descriptions of these models can be found in Chapter 10 of the book (Pinheiro, Bretz, and Branson), "*Dose Finding In Drug Development*" (Springer, 2006) [35]. Figure 2.1 are the graphic presentation of some of these models.

The modeling approach often evaluate only one model at a time as opposed

Figure 2.1: Sample Dose-Response Models



to a candidate set of models simultaneously in one procedure. The successful application of the model-based method in a given dose–response study would depend on the accuracy assumptions of a particular dose–response relationship imposed on the data. The validity of the statistical results could be called into questions under the circumstances where the true underlying dose–response relationship is unknown and is distant from the imposed model. As a matter of fact, the model uncertainty is one of the major pitfalls when using statistical modeling. The intrinsic problem is the introduction of a new source of variability by selecting a particular model M at any stage in the drug dose–response research prior to the final analysis. A common approach to deal with this is to use the information criteria based on a reasonable discrepancy measure to assess the lack of fit so that a set of models can be evaluated. Many model selection criteria are available and the discussion on the best method is always the topic of interest (Zucchini, 2000; Kadane and Lazar, 2004) [39].

A well-known information criterion is the penalized log-likelihood (Akaike,

1973) [2]:

$$AIC = -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + 2p, \tag{2.13}$$

where $\mathcal{L}$ denotes the likelihood function under the fitted model and p refers to the number of corresponding parameters in the model. In the OLS cases, the (2.13) can be written as:

$$AIC = n \log \left(\frac{RSS}{n}\right) + 2p,$$

where RSS is the estimated residual sum of squares. The application of AIC is to calculate the value from (2.13) for each model on the same data set. The best model is the one with the minimum AIC value.

An alternative information criterion in the same general form is known as Bayesian information criterion (BIC) (Schwarz, 1978) [31]:

$$BIC = -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + p\log(n),$$

which also depends on the sample size n. Although both criteria are derived in completely different ways, the BIC differs from the AIC only in the second term, favoring simpler models than AIC as n increases. The reality is that any measure of a fit, either the AIC or BIC or any other criterion, bears the inherent drawback of a missing error control. If we simply select the model corresponding to the best AIC, the validity of model conclusion still could not be confirmed even we choose a candidate set of models for evaluation since the application of the AIC would always lead to one single model to be selected regardless of the goodness of fit given to the observed data. Many statistical inference methods were proposed to deal with this issue.

A different idea is to consider model selection process as a multiple hypotheses testing problem, where the selection of a specific model is performed while controlling the Type I error rate or FWER at a pre-specified level (Shimodaira, 1998; Junquera, et al., 2002)[33]. In a nutshell, this approach would start with a reference set of models and only removes those models shown less significant. This approach may retain more than one model at the end.

2.3 MCP–Mod: Combining multiple comparisons and modeling

F. Bretz, J. C. Pinheiro, and M. Branson (2005) [11] proposed a general framework, MCP–Mod, by integrating the multiple comparisons procedure (MCP) described in Section 2.1 and the modeling (Mod) technique outlined in Section 2.2 into the analysis for the normally distributed dose–response data. J. C. Pinheiro, B. Bornkamp, E. Glimm and F. Bretz (2013)[29] further extended the MCP–Mod to cover a broad range of dose–response data types such as the binary, the counts, or the time-to-event as well as the repeated dose–response measurements over time. This extension named as generalized MCP–Mod (gMCP–Mod), is an overarching framework that covers the normal and the non-normal distributed dose–response data and has been implemented in the R package, **DoseFinding** [9]. In general, the framework starts with defining a candidate set of dose–response models. The underlying dose–response shape (relationship) would be established through an optimal contrasts test (similar to the concept of the trend tests described in Section 2.1) using a multiple comparisons procedure while the Type I error rate is controlled. The "best" model among the significant models identified through the optimal contrasts test would then be evaluated via a chosen information criteria such as the AIC, the generalized AIC (gAIC) or the BIC. Finally the best model is used to predict an optimal dose in the sense of obtaining the MED.

2.3.1 MCP–Mod

The MCP–Mod assumes that the dose–response data follows the independent and normal distribution with covariates enter linearly in a standard form:

$$y_{ij} = f(d_i, \boldsymbol{\theta}, \boldsymbol{x}_{ij}) + \varepsilon_{ij} = \boldsymbol{x}_{ij}' \boldsymbol{\theta}_0 + \theta_1 f^0(d_i, \boldsymbol{\theta}^0) + \varepsilon_{ij}, \tag{2.14}$$

where $\boldsymbol{x}_{ij}$ denotes the covariate vector for j^{th} patient in dose group i and $\boldsymbol{\theta}_0$ refers to the vector of linear parameters. Other quantities are the same as defined in (2.12). The MCP–Mod approach is implemented in the following five steps:

Step 1

Assume that there exists a set of M candidate models: $\boldsymbol{M} = \{M_\ell, \ell = 1, \ldots, M\}$. Each of these models can be represented by a fixed mean vector $\boldsymbol{\beta}_\ell^0 = (\beta_{\ell 0}^0, \beta_{\ell 1}^0, \ldots, \beta_{\ell k}^0)'$, which is the sub-vector of the parameter vector $\boldsymbol{\beta}$ as derived from the standardized model $f_\ell^0(d_i, \boldsymbol{\theta}_\ell^0) = \beta_{\ell i}^0, i = 0, \ldots, k$. Note that $\boldsymbol{M}$ may include different models, i.e., $f_\ell^0 \neq f_{\ell'}^0$, or different parameter specifications $\boldsymbol{\theta}_\ell^0 \neq \boldsymbol{\theta}_{\ell'}^0$ for the same model $f_\ell^0 = f_{\ell'}^0, 1 \leq \ell \neq \ell' \leq M$.

Step 2

After the candidate model set $\boldsymbol{M}$ has been identified, the goal is to select the best fitting model(s) while controlling the Type I error rate. To this end, we test the null hypothesis $H_0 : \boldsymbol{c}'\boldsymbol{\beta}^0 = 0$ against the one-sided alternative $H_a : \boldsymbol{c}'\boldsymbol{\beta}^0 > 0$ for a given $(k+1) \times 1$ contrast vector $\boldsymbol{c} = (c_0, c_1, \ldots, c_k)'$ of known constants subject to $\sum_{i=0}^{k} c_i = 0$. The modeling context considered here is that of an ANCOVA model in which the doses are represented by the indicator functions. In the model (2.14), the regression matrix $\boldsymbol{X}$ has rows given by x_{ij} concatenated with a vector of length k with 1 in the ith position and 0 in the remaining (indicating that the patient is in the dose group i). This leads to the construction of a single contrast test:

$$t = \frac{c'\hat{\boldsymbol{\beta}}^0}{\hat{\sigma}\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})_0^- \boldsymbol{c}}},$$

where $(\boldsymbol{X}'\boldsymbol{X})_0^-$ is the sub-matrix of $(\boldsymbol{X}'\boldsymbol{X})^-$ associated with $\boldsymbol{\beta}^{\mathbf{0}}$. Under the assumptions of general linear model (2.6) and the null hypotheses $H_0 : \boldsymbol{c}'\boldsymbol{\beta}^0 = 0$, the test statistic t follows a central t distribution with v degrees of freedom. If H_0 is not true, t follows a non-central t distribution with the non-centrality parameter:

$$\tau = \frac{c'\boldsymbol{\beta}^0}{\sigma\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})_0^-\boldsymbol{c}}}.$$

Note that, at the planning stage of a dose–response study, the parameters $\boldsymbol{\beta}^0$ and σ are assumed known ($\boldsymbol{\beta}^0$ is known through the "guesstimates"), thus, for a fixed design $\mathbf{X}$, the non-centrality parameter τ depends only on the contrast vector $\boldsymbol{c}$. Consequently, to maximize the chance of rejecting $H : \mathbf{c}'\boldsymbol{\beta}^0 = 0$ for a given model, we select the contrast vector $\boldsymbol{c}$ such that the $\tau = \tau(\mathbf{c})$ is maximized. Using the Lagrange's multiplier, one can show that for $\mathbf{S} = (\mathbf{X}'\mathbf{X})_0^-$, the choice of the optimal $\mathbf{c}$ vector:

$$c_{\text{opt}} = \mathbf{S}^- \left(\boldsymbol{\beta}^0 - \frac{\boldsymbol{\beta}^{0'}\mathbf{S}^-\mathbf{1}}{\mathbf{1}'\mathbf{S}^-\mathbf{1}}\mathbf{1}\right) \tag{2.15}$$

maximizes $\tau(\mathbf{c})$ (Bornkamp, 2006) [8]. The solution is unique if $\parallel \mathbf{c}_{opt} \parallel = 1$ is imposed. In the simplest case of the one-way ANOVA model (2.1), the standardized parameter vector $\boldsymbol{\beta}^0$ reduces to the standardized mean vector $\boldsymbol{\mu}^0 = (\mu_0^0, \ldots, \mu_k^0)$ and the covariance matrix $\boldsymbol{S}^- = \text{diag}(n_0^{-1}, \ldots, n_k^{-1})$. Thus the (2.15) simplifies to

$$c_{\text{opt}} = \begin{pmatrix} n_0(\mu_0^0 - \bar{\mu}^0) \\ \vdots \\ n_k(\mu_k^0 - \bar{\mu}^0) \end{pmatrix},$$

where $\bar{\mu}^0 = \sum_{i=0}^{k} n_i\mu_i^0 / \sum_{i=0}^{k} n_i$ is the overall standardized mean value. In this case, the optimal contrast coefficients are simply computed by the group sample sizes n_i and the expected mean responses μ_i^0. In the balanced case (equal sample size for all dose groups), the optimal contrast coefficients do not even depend on the samples sizes and the computation is further simplified.

Step 3

For each model in the candidate set $\boldsymbol{M}$, an optimal contrast vector $\boldsymbol{c}_{opt,\ell}$ is computed from (2.15). These optimal contrasts can then be used to test for a significant dose–response effect model by simultaneously inferencing the parameters $\boldsymbol{c}'_{\text{opt},\ell}\boldsymbol{\beta}^0$. Let the null hypothesis be $H_\ell : \boldsymbol{c}'_{\text{opt},\ell}\boldsymbol{\beta}^0 = 0$, then the associated test statistic is

$$t_\ell = \frac{\mathbf{c}'_{\text{opt},\ell}\hat{\boldsymbol{\beta}}^0_\ell}{\hat{\sigma}\sqrt{\mathbf{c}'_{\text{opt},\ell}(\mathbf{X}'\mathbf{X})^-_0\mathbf{c}_{\text{opt},\ell}}},\ \ell = 1, \ldots, M. \tag{2.16}$$

Consider $t_{\max} = \max_\ell t_\ell$, as discussed in Section 2.1, a way to combine the best statistics t_ℓ into a single decision rule. By constructing of the test for the individual hypothesis H_ℓ, we look for an appropriate critical value q such that $t_{\max} > q$ for concluding a significant dose–response relationship (model). Equivalently, we can assess the multiplicity-adjusted p-value, p_ℓ of t_ℓ such that $p_{\min} = \min_\ell p_\ell < \alpha$.

As discussed before, the vector of t statistics, $\boldsymbol{t}' = (t_1, \ldots, t_M)$, is M-variate t distributed with v degrees of freedom and the correlation matrix $\mathbf{R} = \mathbf{D}\mathbf{C}'(\mathbf{X}'\mathbf{X})^-_0\mathbf{C}\mathbf{D}$. In this correlation matrix $\mathbf{R}$, $\mathbf{C}_{k\times M} = (\mathbf{c}_{\text{opt},1}, \ldots, \mathbf{c}_{\text{opt},M})$ and $\mathbf{D} = \text{diag}(\mathbf{c}'_{\text{opt},\ell}(\mathbf{X}'\mathbf{X})^-_0\mathbf{c}_{\text{opt},\ell})^{-1/2}_\ell$. Then the associated critical value q is $q = t_\alpha(M, v, \mathbf{R})$.

Step 4

Once a significant model has been identified through previous steps, i.e., there exits $t_{\max} > q$, the MCP–Mod uses the single model for the final dose estimation. If more than one model is significant, then all significant models are in a reference set M^*. One can then select a single model by evaluating the standard model selection criteria (AIC or BIC, etc.) values for the final target dose estimation stage. Alternatively, once can apply the model averaging techniques (Buckland et al., 1997) [12] to the M^* for the final model selection.

Step 5

The final step consists of fitting the selected dose–response model to the study data and estimate the target doses(s) of interest using modeling techniques. For estimating the MED, the definitions (2.8) or (2.9) discussed in

Section 2.1.2 are extended to account for the continuous dose range $(d_0, d_k]$. The MED associated with a model $f(d, \boldsymbol{\theta})$ is then defined as the smallest dose in the range $(d_0, d_k]$ for which $f(d, \boldsymbol{\theta}) - f(d_0) \geq \delta$ (absolute scale) or $f(d, \boldsymbol{\theta})/f(d_0) \geq \lambda$ (relative scale), where $\delta > 0$ or $\lambda > 1$ is the clinically relevant effect and $f(d_0, \boldsymbol{\theta})$ is assumed to be positive.

2.3.2 gMCP–Mod

While MCP–Mod is well developed for the homoscedastically and normally distributed dose–response data, in practice, there are needs for such a method being implemented to evaluate the dose–response signals from the binary, counts or time-to-event endpoints and etc. Pinheiro et al. (2013) [29] developed an extension of the MCP–Mod by testing and modeling these type of dose–response data in the context of general parametric models and named this approach gMCP–Mod (generalized MCP–Mod). The key contribution to this extended version of the MCP–Mod is to separate modeling the dose–response signals from estimating the corresponding mean dose–responses. In another word, the mean dose–responses are estimated based on the relevant parameters in their probability distribution first and the dose–response relationships are modeled via these estimated means after. The reason for this separation is that directly maximizing the likelihood (ML) for estimating the model parameters requires derivations of the likelihood in every specific case and would take considerable amount of model-specific coding effort. It become practically and computationally inefficient.

Recall that the optimal contrasts only depend on the standardized dose–response model functions rather than the complete model functions, as shown in (2.15). The parameter vectors $\boldsymbol{\theta}^0$'s are specified in the planning stage via "guesstimates". This provides an opportunity to implement an alternative two-stage modeling to fit the observations using the generalized least squares (GLS) during the MCP testing. Corresponding to the test (2.16) in the MCP–Mod, the test statistic used in the gMCP–Mod is an asymptotic z-test statistic for the hypotheses $H_0 : \mathbf{c}'_{\text{opt},\,\ell}\,\boldsymbol{\mu}_\ell = 0$ vs. $H_a : \mathbf{c}'_{\text{opt},\,\ell}\,\boldsymbol{\mu}_\ell > 0$:

$$z_\ell = \frac{\mathbf{c}'_{\text{opt},\,\ell}\,\hat{\boldsymbol{\mu}}}{\sqrt{\left[\mathbf{C}'_{\text{opt}}\,\hat{\mathbf{S}}\mathbf{C}_{\text{opt}}\right]_{\ell,\ell}}},\ \ell = 1, \ldots, M, \tag{2.17}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{S}}$ are the ANOVA estimates using GLS in the first stage, $\mathbf{C}_{\text{opt}} = (\mathbf{c}_{\text{opt},\,1}, \ldots, \mathbf{c}_{\text{opt},\,M})$ is the optimal contrasts matrix pre-specified using the "guesstimates" and $[\ ]_{\ell,\ell}$ is the ℓ^{th} diagonal element of matrix $[\ ]$. It is approved that $\mathbf{C}'_{\text{opt}}\,\hat{\boldsymbol{\mu}}$ asymptotically follows the normal distribution with the mean $\mathbf{C}'_{\text{opt}}\,\boldsymbol{\mu}$ and the estimated covariance matrix $\mathbf{C}'_{\text{opt}}\,\hat{\mathbf{S}}\mathbf{C}_{\text{opt}}(M \times M$ in dimension). Note that $\mathbf{C}_{opt}$ is a k(doses)$\times M$ matrix.

The test statistic used for establishing an overall dose–response signal is the $z_{\max} = \max\limits_{1<\ell\leq M} z_\ell$. The critical values for the tests with (asymptotically) exact level α can be derived from the joint distribution of $z = (z_1, \ldots, z_M)$, which can be obtained from the joint distribution of the contrast estimates given previously by using

$$P(z_{\max} > q) = 1 - P(z_{\max} \leq q) = 1 - P(z \leq q\mathbf{1}).$$

The quantity $q\mathbf{1}$ is an integral over the multivariate normal distribution, which can be evaluated using numerical integration. The R package, **mvtnorm**, includes a function to numerically calculate the quartiles and also the probabilities of the underlying multivariate normal distribution (Genz A, Bretz F., 2009) [18].

If more than one model is identified, a model selection criterion in the gMCP–Mod is used for the model fitting comparison. The criterion named gAIC is defined as

$$\hat{\boldsymbol{\Psi}}(\hat{\boldsymbol{\theta}}) + \dim(\boldsymbol{\theta})\tau,\ \text{where } \hat{\boldsymbol{\Psi}} = (\hat{\boldsymbol{\mu}} - f(\mathbf{x}, \hat{\boldsymbol{\theta}}))'\widehat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{\mu}} - f(\mathbf{x}, \hat{\boldsymbol{\theta}}))$$

(Pinheiro et al, 2013) [29]. In practice, $\tau = 2$ is used to calculate the gAIC value. Refer to the same article [29] for detailed discussions on the gAIC.

CHAPTER 3

RESAMPLING AND HYPOTHESIS TESTING

In the end of Section 2.2, we briefly touched upon the concept of considering the model selection process as a multiple hypotheses testing problem. The MCP–Mod or the gMCP–Mod described in Section 2.3 actually take the concept into a framework that evaluates the multiple dose–response models from the multiple hypotheses perspective via an optimal contrast trend testing procedure. In this chapter, we introduce the resampling-based statistical methods and extend our review to the permutation test as well as the relevant resampling-based multiple testing procedures that will be used in our proposed approach to be discussed in Chapter 4.

3.1 Resampling Statistical Methods

Resampling methods refers to the statistical methods based on available observations (observed samples) rather than a set of standard assumptions of the underlying populations. They represent a family of distinctly different statistical analyses from that of the traditional statistics. The common resampling statistical methods include jackknife, bootstrap, cross-validation and permutation tests.

Jackknifing is used in the statistical inference to estimate the bias and standard error (variance) of a statistic that is calculated based on a random sample of the original observations. It was invented by Quenouille in 1949 and extended by Tukey in 1958. Quenouille invented this method with the intention of reducing the bias of the sample estimate. Tukey extended it by assuming that if the replicates could be considered independently and identically distributed, then an estimate of the variance of a sample parameter could be made and that it would be approximately distributed as a t variate with $n - 1$ degrees of freedom with n being the sample size. The basic idea behind the jackknife variance estimator lies in systematically recomputing the statistic estimate, leaving out one or more observations at a time from the sample set. From the new set of replicates of the statistic, the estimate for the bias or for the variance of the statistic can be calculated.

Bootstrapping is a method for assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measures) to sample estimates. This technique allows the estimations of the sampling distribution of almost any statistic using only very simple methods. It is the practice of estimating properties of an estimator when sampling from an approximating distribution. One standard choice for an approximate distribution is the empirical distribution of the observed data. In the case where a set of observations can be assumed to be from an independently and identically distributed population, one can implement the bootstrapping by constructing a number of resamples from the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling *with replacement* from the original dataset. It may also be used for constructing hypothesis tests. The bootstrap method is often used as an alternative to the inference based on parametric assumptions when those assumptions are in doubt, or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors. The bootstrap is also seen in the applications for cross validating statistical models.

Cross-validation is a resampling statistical method for validating a predictive model. The subsets of the data are held out for the use as the validating sets. A model is fit to the remaining data (a training set) and used to predict for the validation set. Averaging the quality of the predictions across the validating sets yields an overall measure of prediction accuracy. One form of the cross-validation leaves out a single observation at a time; this is similar to the jackknife. Another, K-fold cross-validation, splits the data into K subsets; each is held out in turn as the validation set. This avoids “self-influence”. For comparison, in the regression analysis methods such as the linear regression, each y value draws the regression line toward itself, making the prediction of that value appear more accurate than it really is. Cross-validation applied to the linear regression predicts the y value for each observation without using that observation.

Permutation test (also called randomization test, re-randomization test, or exact test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of that test statistic under the rearrangements of the labels (identifier) on the observed data points. We will devote the next section to detailed discussions since the permutation test is directly relevant to our proposed approach in this dissertation.

3.2 Permutation Tests

In a decision based on a statistical significance testing, we usually start with formulating the problem into hypotheses of interest. We subsequently collect data by conducting experiments or searching for available information. In analyzing the observations, a test statistic is chosen and the value of the test statistic is computed based on the observed samples. We then compare the observed statistic value against the distribution of the test statistic under the null hypothesis to guide us for making statistical decisions. For parametric statistics, the distribution of the data is assumed to be known, the limiting distribution of a test statistic under the null hypothesis can be specifically determined according to the central limit theorem. For the non-parametric statistics, the distribution of the data is often unknown and the distribution of a test statistic is often constructed using reampling methods. The permutation test is one of the non-parametric tests in this perspective.

Permutation testing can be traced back to Fisher in 1935 (Fisher's exact test is an example of a commonly used permutation test for evaluating the association between two dichotomous variables). Since then, the practical applications of such methods have increased steadily with the grow of computing power. In any permutation tests, instead of comparing the actual value of a test statistic to a standard statistical distribution, the referential distribution is generated from the data themselves by the permutation resampling. The permutation method provides an efficient approach to testing when the data do not conform to the distributional assumptions of the statistical method one wants to use (e.g. normality).

Permutation Distribution for Test Statistic

Suppose that a new treatment for a specified disease is being compared to a control by some observed clinical effect on the patients from each group. Of the N patients available for the study, n are randomly assigned to receive the new treatment, while the remaining m $(N - n)$ receive the control. The null and alternative hypotheses of interest are:

H_0 : There is no difference between the new treatment and the control,

H_a : The new treatment is better than the control.

Denote the endpoint of the clinical effect for the new treatment and the control by $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_m$, respectively. To measure the difference in the endpoint between the two groups, we might calculate the difference in mean as $T = \bar{Y} - \bar{X}$. We wish to determine if this difference in mean is extreme enough in some reference distribution to suggest that the new treatment is better.

If the null hypothesis is true, i.e., there is no difference between the two groups, then the clinical effect on each patient will be the same *regardless of which group the study patient is assigned to.* Hence, the basis for building a probability distribution for $\bar{Y} - \bar{X}$ (considered as the test statistic for this comparison) under the null hypothesis only comes from the randomization of the available patients to the two groups, i.e., n and m patients being randomly assigned to the new treatment and the control respectively. The observed sample is just one of $\binom{N}{n}$ equally likely permutations that could have occurred. If we calculate $T = \bar{Y} - \bar{X}$ for each of the possible permutations, the probability distribution for T, called permutation distribution, is constructed.

As we can see, in comparing group differences, the permutation resamples must be drawn in a way that is consistent with the null hypothesis and with the study design. They are drawn at random from the original data *without replacement*, in contrast to the bootstrap samples, which are drawn *with replacement*.

p-value in Permutation Test

The p-value of a permutation test for an H_0 can be calculated as the probability of getting a test statistic as extreme as, or more extreme than (in favor of the H_a), the observed test statistic t_{obs} (not necessary a t statistic). Since all of the $\binom{N}{n}$ random permutations are equally likely under the H_0. The p-value of the significant test for a location parameter between the two independent groups can be computationally define as:

$$p = \Pr\left(T \geq t_{\text{obs}} \mid H_0\right) = \frac{\sum_{i=1}^{\binom{N}{n}} I(t_i \geq t_{\text{obs}})}{\binom{N}{n}}, \tag{3.1}$$

where t_i is the value of the test statistic $T = \bar{Y} - \bar{X}$ for the ith permutation sample and $I(\cdot)$ is the indicator function.

It is clear from the discreteness of the permutation distribution that the p-value must be a multiple of $1/\binom{N}{n}$. If we choose a significance level $\alpha = k/\binom{N}{n}$, one of the achievable p-values, then it becomes

$$\begin{aligned} \Pr\text{ (Type I error)} &= \Pr\ (P \geq \alpha | H_0) \\ &= \Pr\left(\sum_{i=1}^{\binom{N}{n}} I(t_i \geq t_{obs} | H_0\right) \\ &= \frac{k}{\binom{N}{n}} = \alpha, \end{aligned} \tag{3.2}$$

where P is a random variable of p-values. This formula indicates that the permutation test of the H_0 is an exact test. If α is not chosen as one of the achievable p-values but $k/\binom{N}{n}$ is the largest achievable p-value less than α, then Pr (Type I error) $= k/\binom{N}{n} < \alpha$ and the permutation test is conservative. Either way, the test is guaranteed to control the probability of a Type I error under very minimal conditions: permutation of the subjects to treatments.

Monte Carlo Sampling

Constructing the permutation distribution for a test statistic involves careful enumeration of all $\binom{N}{n}$ divisions of the observations. This poses two computational challenges. First, the sheer number of required calculations could become very large even the sample size is only moderate. There are over 155 million ways to split 30 observations into two groups of size 15, i.e., $\binom{30}{15} = 155,117,520$, and over 5.5 trillion ways to divide them into three groups of size 10. Second, enumerating each unique division of the data is not easily programmed and requires specialized software such as the **coin** package in R.

One easy, and very practical solution to both these challenges is to use Monte Carlo sampling from a theoretical permutation distribution to construct a *sampling permutation distribution* of the test statistic to estimate the exact p-value. Since the p-value is simply the proportion of the test statistic values from the permutation samples as extreme or more extreme than the observed test statistic value, we can naturally estimate this by randomly choosing the samples from all of $\binom{N}{n}$ permutations and obtain the test statistic value for each random permutation sample and calculating a sample proportion that is as extreme or more extreme than the observed value. This can be easily accomplished by repeatedly and randomly dividing the N observations into groups of size m and n (in above example) and calculating the test statistics. A few thousands of the test statistic values usually are sufficient to get an accurate estimate of the exact p-value. If the number of B test statistic values, t_i, $i = 1, \ldots, B$, is randomly sampled (calculated), a one-sided Monte Carlo p-value for the test that rejects H_0 for the large values of t is

$$\hat{p} = \frac{1 + \sum_{i=1}^{B} I(t_i \geq t_{\text{obs}})}{B+1}. \tag{3.3}$$

Including the observed value t_{obs}, there is a total of $B+1$ realizations of the test statistic. Since the observed value will always be "as extreme" as itself if the H_0 is not true, the Monte Carlo p-value will be no smaller than $1/(B+1)$. This is consistent with the exact p-value, which must be at least $1/\binom{N}{n}$. The idea of sampling from the permutation distribution was first proposed by Dwass (1957) [15]. The test remains exact and conditionally distribution free with the only penalty being a small loss of efficiency. Jöckel (1986) [23] showed that this loss of efficiency decreases as B increases and he gave a lower bound for the loss of efficiency as a function of B.

We now illustrate the permutation test using the example in Chapter 18 [21] of the book "*The Practice of Business Statistics*": Do reading activities increase Degree of Reading Power (DRP) scores? The study assigned students at random to either a new method (treatment group, 21 students) or the traditional teaching methods (control group, 23 students). Their DRP scores at the end of the study appear in Table 3.1. The statistic measuring whether the new method is successful is the difference in mean DRP scores,

$$T = \bar{X}_{\text{treatment}} - \bar{X}_{\text{control}}.$$

The null hypothesis is that there is no difference between the two methods. If this is true, the DRP scores in Table 3.1 should not depend on the teaching

Table 3.1: DRP scores for third-graders

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

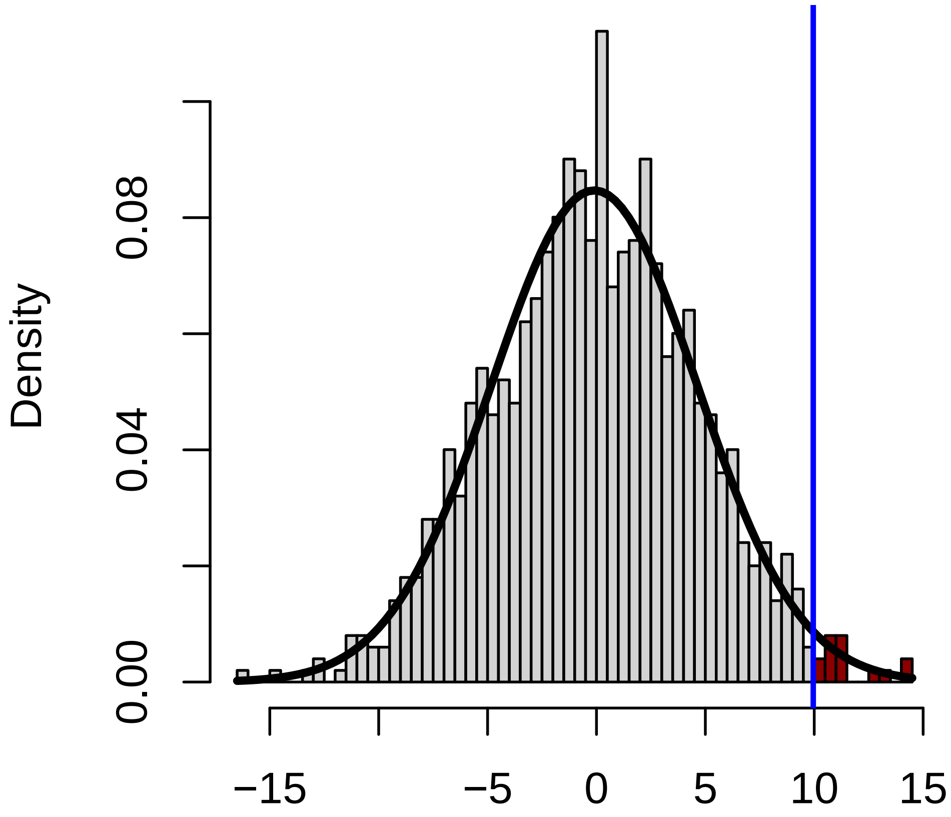
methods and the observed difference in the group means just reflects the accident of a random assignment to the two groups. One can make many repetitions of the random assignments with each student always keeping his or her original DRP score unchanged, i.e., draw samples from these scores but scramble the student assignments to the study groups. For each resample, the difference in DRP between the group means can be calculated. A test based on these resampled group mean differences is called the permutation test, from the mathematical name for scrambling a set of items (treatment assignments).

Here is an outline of the permutation test procedure for comparing the mean DRP scores:

- Choose 21 of the 44 students at random to be in the treatment group; the other 23 are in the control group. This is an ordinary simple random sample, chosen *without replacement* as a permutation resample. Calculate the mean DRP score in each group, using the individual scores in Table 3.1. The difference between these two means is our statistic for this resample.

- Repeat this resampling hundreds of times from the 44 students and calculate the mean DRP scores and the difference each time. The distribution of the statistic from these resamples forms a sampling distribution under the condition that the null hypothesis that H_0 is true. It is called a

Figure 3.1: The permutation distribution of the statistic T based on 999 permutation resamples from the DRP scores of 44 students. The observed $t_{\text{obs}} = 9.954$, vertical line, is in the right tail



sampling permutation distribution, which comes from Monte Carlo sampling. Note that, in practice the complete permutation distribution is barely constructed for the reason that any moderate sample size could lead to a larger number of permutations. Therefore, the term, **permutation distribution**, is often used to refer to the sampling permutation distribution. We will use this convention in the rest of this thesis.

- The value of the statistic actually observed in the study was

$$t_{\text{obs}} = \bar{X}_{\text{treatment}} - \bar{X}_{\text{control}} = 51.476 - 41.522 = 9.954$$

Graphically, one can locate this value on the permutation distribution (dark area to the right of the vertical line as shown in Figure 3.1) to obtain the estimated p-value. The calculation of the estimated p value is given by (3.3):

$$\hat{p} = \frac{1+14}{1+999} = 0.015$$

.

The estimated p-value gave the good evidence that the new method fares better than the traditional reading methods at the significance level of 0.05.

3.3 Resampling-based Multiple Testing Procedures

When such a permutation test described above is extended to compare more than two groups, the multiplicity issue arises as in the traditional statistics. One needs corresponding multiple testing procedures to control the Type I error rate of FWER. As a matter of fact, the resampling-based multiple testing procedures have been developed by many experts. The major advantages of the resampling-based multiple testing procedures are 1) making fewer assumptions about the data-generating process and therefore yielding more robust multiple test procedures; 2) using data-driven distributional characteristics, including discreteness and correlation structures that result in more powerful procedures. In the meantime, the obvious drawbacks of these methods tend to be 1) approximate and requiring large sample sizes; 2) computational intensive and difficult if complexities of data models are involved in the simulations.

The closure principle in the multiple testing problems provides a convenient, flexible and powerful foundation to discuss the resampling-based multiple testing procedures. Let $T_1, \ldots, T_m$ denote the random variables representing the test statistics associated with m hypotheses of interest H: $H_1, \ldots, H_m$ and $t_1, \ldots, t_m$ denote the observed test statistic values. In general, any α-level test may be used to test intersection hypotheses in the closed family H. Popular procedures include the minimum p-value based such as Bonferroni procedure or maximum test statistic such as Dunnett procedure. For different statistical models or different real problem settings, there are different choices of test statistics. For examples, the F-statistics can be used for testing the global intersection hypothesis in an ANOVA model for dose-finding studies; Fisher combination tests may be used in multi-center, subgroup or other analyses that intend to combine data across factors or locations.

Because of intensive computation effort is often involved in the resampling methods when the number of hypotheses m is large, to test $2^m - 1$ intersections in the closed family would become impossible. Many researchers have developed procedures which have dramatically reduced the computational burden based on the following key assumptions:

- For each non-empty index set $I \subseteq \{1, \ldots, m\}$, the intersection hypothesis $H_{0I} = \bigcap_{i \in I} H_{0i}$ is tested using the maximum statistic $T_{\max}(I) = \max_{i \in I} T_i$.
- The *subset pivotality condition* (Westfall and Young, 1993) [38] is present, i.e., for each non-empty index set I, the distribution of $T_{\max}(I)$ under H_{0I} is identical to the distribution of $T_{\max}(I)$ under the global null hypothesis $H_{\mathbf{0}}$.

3.3.1 Single-step resampling-based procedures

Westfall and Young (1993) [38] demonstrated that in a single step manner, the multiplicity-adjusted p-value is the proportion of the total number of minimum p-values over the marginal (observed) p_i corresponding to the null hypothesis H_{0i} calculated from all samples for which these minimum permutated p-values are less than or equal to the particular original observed p-value.

If we know the joint distribution of the p-values, the single-step adjusted

p-values can be computed based on:

$$\Pr\left(\text{Reject at least one} H_{0i} \mid H_{0I}\right) = \Pr\left(\min_{1\leq i\leq m} p_i \leq q \mid H_{\mathbf{0}}\right),$$

where q is the pre-specified critical value for controlling the FWER. The methods to determine the q quantity include well-known Bonferroni method and its related, more powerful modifications. These single-step methods are applicable mainly assuming that the marginal p_i's follow $U[0,1]$ independently under the null hypotheses.

In cases where there are correlations among the marginal p_i's, the joint distribution of $\Pr\left(\min_{1\leq i\leq m} p_i \leq q \mid H_{\mathbf{0}}\right)$ is not identifiable, alternatively to other methods in which some assumption of the correlation is needed, the resampling methods can be used to simulate the joint distribution since the resampling p-value vector has the same distribution as the observed p-value vector under the null hypothesis per *subset pivotality condition* being met. The resampling single-step adjusted p-value can then be defined as:

$$\tilde{p}_i = \Pr\left(\min_{1\leq j\leq m} P_j \leq p_i \mid H_{\mathbf{0}}\right), \tag{3.4}$$

where $\tilde{p}_i$ is probability that the smallest p-value from the random variable P_j on a resample dataset is smaller than the observed p-value p_i from the original dataset. Computationally, it is calculated as the proportion of the minimum p-values from all permutation resamples (the reference distribution, i.e, the permutation distribution of the minimum p-value from all resamples) that falls below the particular observed p-value, p_i.

Alternately, we may consider the procedure based on the single-step maximum T adjusted p-values which are defined in terms of the test statistics T_i:

$$\tilde{p}_i = \Pr\left(\max_{1\leq j\leq m} T_j \geq t_i \mid H_{\mathbf{0}}\right). \tag{3.5}$$

3.3.2 Step-down resampling-based procedures

The single step method is overly conservative in the situation where more of null hypotheses are false. The benefit of the two assumptions listed above lead to that stepwise resampling-based procedures similar to the Holm and stepwise Dunnett procedures can be constructed.

The step-down procedures of a resampling-based multiple testing was developed independently by Westfall and Young (1993) [38], Blair and Karniski (1994) [6] and Troendle (1995) [36]. The general algorithm of the step-down procedure is as follows:

Let $H_{(1)}, \ldots, H_{(m)}$ denote the m hypotheses corresponding to the ordered test statistics $t_{(1)} > \ldots > t_{(m)}$. The step-down resampling-based procedure is defined as follows:

- Step 1. Reject $H_{(1)}$ if
$$\Pr\left(\max(T_1, \ldots, T_m) \geq t_{(1)}\right) \leq \alpha$$
and go to the next step. Otherwise retain $H_{(1)}, \ldots, H_{(m)}$ and stop.
- Step $i = 2, \ldots, m-1$. Reject $H_{(i)}$ if
$$\Pr\left(\max(T_i, \ldots, T_m) \geq t_{(i)}\right) \leq \alpha$$
and go to the next step. Otherwise retain $H_{(i)}, \ldots, H_{(m)}$
- Step m. Reject $H_{(m)}$ and stop.

A corresponding step-down procedure based on the minimum p-values was given by Westfall and Yong (1993) [38]. The adjusted p-values from this procedure are defined sequentially:

$$\begin{aligned}
\tilde{p}_{(1)} &= \Pr\left(\min_{l \in \{r_1, r_2, \cdots, r_k\}} P_l \leq p_{(1)} \mid H_0\right) \\
\tilde{p}_{(2)} &= \max\left[\tilde{p}_{(1)},\ \Pr\left(\min_{l \in \{r_2, \cdots, r_k\}} P_l \leq p_{(2)} \mid H_0\right)\right] \\
&\vdots \\
\tilde{p}_{(j)} &= \max\left[\tilde{p}_{(j-1)}\ \Pr\left(\min_{l \in \{r_j, \cdots, r_k\}} P_l \leq p_{(j)} \mid H_0\right)\right] \\
&\vdots \\
\tilde{p}_{(k)} &= \max\left[\tilde{p}_{(k-1)}\ \Pr(P_{r_k} \leq p_{(k)} \mid H_0)\right],
\end{aligned} \tag{3.6}$$

where $r_1, r_2, \cdots, r_k$ are the indexes for the ordered p-values, $p_{(k)} > \ldots > p_{(1)}$ corresponding to $H_{(k)}, \ldots, H_{(1)}$.

CHAPTER 4

PROPOSED PERMUTATION TEST IN EVALUATION OF DOSE-RESPONSE MODELS

As we described in Section 2.3 of Chapter 2, the “guesstimates” is the necessary information used for optimal contrasts calculation in MCP–Mod/gMCP–Mod framework. Our motivation as indicated in Section 1.1 is to explore alternatives that can avoid the use of “guesstimates” in the model selection process while the multiplicity is still being taken into consideration. Through review of the resampling statistical methods (Chapter 3), we feel that the well developed permutation testing methodology that has been applied in different fields may provide us an opportunity to tackle the issue.

The permutation test in analyzing dose–response measures has not been extensively researched. For the binary dose-response data, B. Klingenberg (2009) [24] proposed a signed likelihood ratio test statistic as the permutation test in the step of identifying significant dose–response models. In our proposal, RMCP-Mod, we will discuss a similar test under the MCP–Mod for the normal data. The proposed permutation test procedure can also be extend to a larger scope of dose–response data types such as binary, counts, survival events (time-to-event) and etc, as the MCP–Mod was to the gMCP–Mod. We will focus on normally distributed dose–response data in this chapter to lay the fundamentals of the RMCP-Mod and move to its extension, gRMCP-Mod, to different types of dose–response data in the next chapter.

4.1 Dose–Response Model and Test Statistic

We assume that there exists a set of candidate models: $M = \{\ell,\ \ell = 1, 2, \ldots, M\}$. Each of these models has the same general response form:

$$y_{ij,\,\ell} = \mu_{i,\,\ell} + \varepsilon_{ij} = f_\ell(d_i,\ \boldsymbol{\theta}_\ell) + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim \text{iid } N(0, \sigma_\ell^2) \tag{4.1}$$

$$i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i,$$

where i represents k doses in a numeric scale with $i = 1$ referring to a control (0 dose level), and $\mu_{i,\,\ell} = f_\ell(d_i,\ \boldsymbol{\theta}_\ell)$ is the unknown treatment mean with mean vector $\boldsymbol{\mu}_\ell = (\mu_1, \ldots, \mu_k)_\ell$ for a family shape of functions parameterized by a vector of $\boldsymbol{\theta}_\ell$. We hope that these models can broadly cover the dose–response space that includes the underlying dose–response relationship in the observed data.

We also define a flat model, which represents the no-dose–response data (dose effect is not a function of dose levels) as:

$$y_{ij,\,0} = f_0(d_i) + \varepsilon_{ij} = \beta + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim \text{iid } N(0, \sigma_0^2),$$

$$i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i.$$

The f_0 is actually a constant treatment mean model in which treatment means are independent of the dose levels d_i.

The question of whether any of these M models would closely express the existed underlying dose–response relationship can be formulated into the following hypotheses with respect to a quantitative measure without loss of generality:

$$H_{0\ell} : f_\ell = f_0 \text{ vs. } H_{1\ell} : f_\ell > f_0, \text{ for at least one } \ell\ (1 \leq \ell \leq M). \tag{4.2}$$

For testing any $H_{0\ell}$, we propose to choose a penalized likelihood ratio test (LRT) as the test statistic since the likelihood ratio test (LRT) has been widely used to compare two nested models. The $H_{0\ell}$ involves M pairwise comparisons between g_ℓ and g_0. Each comparison of $H_{0\ell}$ can be considered a two-nested model comparison in the sense that f_0 is nested in f_ℓ. The penalized LRT is defined as:

$$\text{LRT}_\ell = -2\ln(L_0 - L_\ell) - 2(p_\ell - p_0), \tag{4.3}$$

where L_ℓ and L_0 are the likelihood quantities from the models f_ℓ and f_0 correspondingly; p_ℓ and p_0 are the numbers of unknown parameters in the models f_ℓ and f_0 respectively. Note that $p_0 = 2$ referring to two unknown parameters β and σ_0^2 in our defined flat model. The term $2(p_\ell - p_0)$ is the penalty for extra model parameters in f_ℓ compared to f_0.

The maximum likelihood estimation of these model parameters β, $\boldsymbol{\theta}_\ell$ from normally distributed dose–response data is illustrated below:

$$\begin{aligned}
&-2\ln(L_0 - L_\ell)\\
&= -2\ln\Bigg[\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma_0^2}(y_{ij}-\beta)^2}\\
&\qquad -\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left(\frac{1}{2\pi\sigma_\ell^2}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma_\ell^2}(y_{ij}-f_\ell(d_i,\,\boldsymbol{\theta}_\ell))^2}\Bigg]\\
&= -2\ln\Bigg[\left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\beta)^2\right\}\\
&\qquad -\left(\frac{1}{2\pi\sigma_\ell^2}\right)^{\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_\ell^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-f_\ell(d_i,\ \boldsymbol{\theta}_\ell))^2\right\}\Bigg]\\
&= -2\Bigg[\left(-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma_0^2)-\frac{1}{2\sigma_0^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\beta)^2\right)\\
&\qquad -\left(-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma_\ell^2)-\frac{1}{2\sigma_\ell^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-f_\ell(d_i,\ \boldsymbol{\theta}_\ell))^2\right)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&= \left(n\ln(\sigma_0^2) + \frac{1}{2\sigma_0^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \beta)^2 \right) \\
&\qquad\qquad - n \left(\ln(\sigma_\ell^2) - \frac{1}{2\sigma_\ell^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - f_\ell(d_i,\ \boldsymbol{\theta}_\ell))^2 \right)
\end{aligned}
$$

The test statistic -2 ln $(L_0 - L_\ell)$ is maximized at:

$$
\begin{aligned}
\hat{\beta} &= \frac{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{n} = \bar{Y} \\
\hat{\sigma}_0^2 &= \frac{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{Y})^2}{n} \\
\hat{\boldsymbol{\theta}}_\ell &= \underset{\boldsymbol{\theta_\ell}}{\text{argmax}} \left[\frac{1}{n}\ln\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_m(d_i,\ \boldsymbol{\theta}_\ell) \right] \\
\hat{\sigma}_\ell^2 &= \frac{\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(y_{ij} - f_\ell(d_i,\ \hat{\boldsymbol{\theta}}_\ell)\Big)^2}{n}
\end{aligned}
$$

The value of the statistic given by a sample of observations will be:

$$
\begin{aligned}
\text{LRT}_m &= -2(L_0 - L_\ell) - 2(p_\ell - p_0) \\
&= \big[n\ln(\hat{\sigma}_0^2) + 1\big] - \left[n\ln(\hat{\sigma}_\ell^2) - \frac{1}{\hat{\sigma}_\ell^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(y_{ij} - f_\ell(d_i,\ \hat{\boldsymbol{\theta}}_\ell)\Big)^2 \right] \\
&\quad - 2(p_\ell - p_0) \\
&= (\text{AIC}_0 - \text{AIC}_\ell) \qquad (4.4)
\end{aligned}
$$

As the above formula shows that this LRT test statistic is actually the difference of Akaike information criteria (AIC) between f_ℓ and f_0, measuring the goodness of fit of the candidate model f_ℓ to the data compared to the no-dose-response model f_0. The larger value of the LRT would be in favor of f_ℓ to f_0. Common statistical softwares can readily compute the AIC, therefore the LRT.

4.2 Computational Procedure

One of the characteristics of permutation tests is the computational intensity due to their resampling nature. We will dedicate this section to describe the computational procedure needed for our proposed permutation test.

Similar to the MCP–Mod/gMCP–Mod framework, our approach also starts with defining a candidate set of models. What follows next is that, instead of using the "guesstimates" to specify model parameters $\boldsymbol{\theta}_\ell$ in (4.1) as in the MCP–Mod/gMCP–Mod, models in the candidate set as well as the comparator, the flat model f_0, are directly fitted to the original experiment dose–response observations from the study. The observed value of the LRT, denoted as $\text{LRT}_\ell^{(\text{obs})}$, can then be obtained from the model fitting results.

Next, we continue to fit these models to a large number (B) of permutation samples drawn from the original dataset without replacement. Under the null hypothesis $H_{0\ell} : f_\ell = f_0$, dosages of $d_1, d_2, \ldots, d_k$ are completely exchangeable. This implies that subjects in an experiment study could be randomly rearranged to the different dose levels through the permutation resampling if the null hypothesis is true. With this exchangeability, which is the key to the permutation test, the total number of the rearrangements in a complete permutation experiment would be:

$$\frac{\left(\displaystyle\sum_{i=1}^{k} n_i\right)!}{\prod_{i=1}^{k} n_i!}.$$

If a pool of sample size $N = \sum_{i=1}^{k} n_i$ is large, building the exact permutation distribution for a test statistic is obviously impractical. Therefore, the Monte Carlo sampling (Section 3.2) is used to construct an approximate distribution of the test statistic by repeating the permutation resampling for B times. B may be "in million preferable" (Westfall and Troendle, 2008) [20], which still could be a small fraction of $(\sum_{i=1}^{k} n_i)!/\prod_{i=1}^{k} n_i!$.

In performing these permutation draws, we technically shuffle the observed responses y_{ij} randomly without replacement and "reattach" the experimenters' original dosing labels (the original dosing assignments to the subjects) to the shuffled responses y_{ij}. The process of the shuffling and reattaching is repeated for B times to form a Monte Carlo sample. During the repeatings, the LRTs (4.4) is calculated for bth permutation sample, $b = 1, 2, \ldots, B$, for each model and it is denoted as:

$$\mathrm{LRT}_{\ell}^{(b)} = \mathrm{AIC}_0 - \mathrm{AIC}_{\ell}^{(b)} \tag{4.5}$$

Note that AIC_0 needs not to be calculated from the permutation samples since it is basically the likelihood from the model f_0 (fitting the flat model without dose labels being involved). AIC_0 takes the same value as it does from the original dataset. After completion of above computations, the permutation distribution (a Monte Carlo distribution) of LRT_{ℓ} is actually constructed for each model, $\mathrm{LRT}_1, \mathrm{LRT}_2, \ldots, \mathrm{LRT}_M$, through the corresponding B number of LRTs.

The subsequent step is to calculate the significance or the observed p-values based on the observed $\mathrm{LRT}_{\ell}^{(obs)}$ for each model by comparing the $\mathrm{LRT}_{\ell}^{(obs)}$ to the corresponding permutation distribution of LRT_{ℓ}. This is done via following formula:

$$p_{\ell}^{(obs)} = \left(\frac{1}{B+1}\right)\left[1 + \sum_{b=1}^{B} I(\mathrm{LRT}_{\ell}^{(b)} \geq \mathrm{LRT}_{\ell}^{(obs)})\right], \tag{4.6}$$

where B is the total number of permutations and $I(\cdot)$ is the indicator function that takes 1 when the comparison in the parentheses is true and 0 otherwise. This is practically a calculation of the quantile of the area where $\mathrm{LRT}_{\ell} > \mathrm{LRT}_{\ell}^{(obs)}$ on the distribution of LRT_{ℓ}. For each permutation sample, we will also obtain the p-values for the candidate models from the corresponding distributions of $\mathrm{LRT}_1, \mathrm{LRT}_2, \ldots, \mathrm{LRT}_M$. The calculation of the bth permutation p-value (the p-value based on permutation sample data) for model f_m is denoted by:

$$p_{\ell}^{(b)} = \left(\frac{1}{B+1}\right)\left[1 + \sum_{i=1}^{B} I(\mathrm{LRT}_{\ell}^{(i)} \geq \mathrm{LRT}_{\ell}^{(b)})\right], \tag{4.7}$$

Through calculating the permutation p-value for every model in the candidate set, the minimum p-value corresponding to the maximum LRT for each permutation sample can be obtained. The bth minimum p-value is chosen over all model p-values:

$$p_{min}^{(b)} = \min(p_1^{(b)}, p_2^{(b)}, \ldots, p_M^{(b)}). \tag{4.8}$$

After these calculations being repeated for B times, the distribution of the minimum p-value is actually constructed. Recall that our hypotheses defined at the beginning is a multiple testing problem if $M > 2$. We can now use the permutation distribution of the minimum p-values to adjust the observed or marginal p-values $p_\ell^{(obs)}$ using the concept described in Section 3.3. A critical value c can be chosen to an α percentile of the distribution and compared to the observed p-value (4.6) on the same distribution for a single-step multiple comparisons adjustment. The single-step adjusted p-value is then defined as

$$\tilde{p}_\ell = \Pr\left(\min_{1\leq j\leq m} P_j \leq p_\ell^{(obs)} \mid H_0\right), \tag{4.9}$$

where $\tilde{p}_\ell$ is the probability for model f_ℓ that the random variable of the smallest p-values from the permutation samples are smaller than observed p-value $p_\ell^{(obs)}$. The computational formula is as below:

$$\tilde{p}_\ell = \left(\frac{1}{B+1}\right)\left[1 + \sum_{b=1}^{B} I\left(p_{min}^{(b)} \leq p_\ell^{(obs)}\right)\right]. \tag{4.10}$$

4.3 Hypothetical and Real Examples

In this section, we make an effort to illustrate the proposed permutation test applying to four examples. The first example is a hypothetical study with the data being generated for an apparent dose–response pattern. The second example is also a hypothetical study in which the data was generated with seemingly weaker dose–response signal. The third example is a real dose ranging study with its summary data available in a public domain. The fourth example is also a real clinical trial with its dose–response information at patient level being used. Through displaying the results from these examples, we would like to show that RMCP-Mod is a viable computational alternative to the model selection process in the MCP–Mod methodology without “guesstimates” being used.

4.3.1 Data with Apparent Dose–Response

A dose–response study was simulated from the normal distributions of different means for 6 independent dosage groups with the same sample size of 30. The summary statistics of the simulated dose responses are shown in Table 4.1 and plotted in Figure 4.1. Note that the assumption of the equal variance for different dosing groups is not necessary when a permutation test is used. This is because that in general, the permutation tests can apply to the observations from different distributions, normal or not, so long as they are symmetric or almost symmetrical about location parameters.

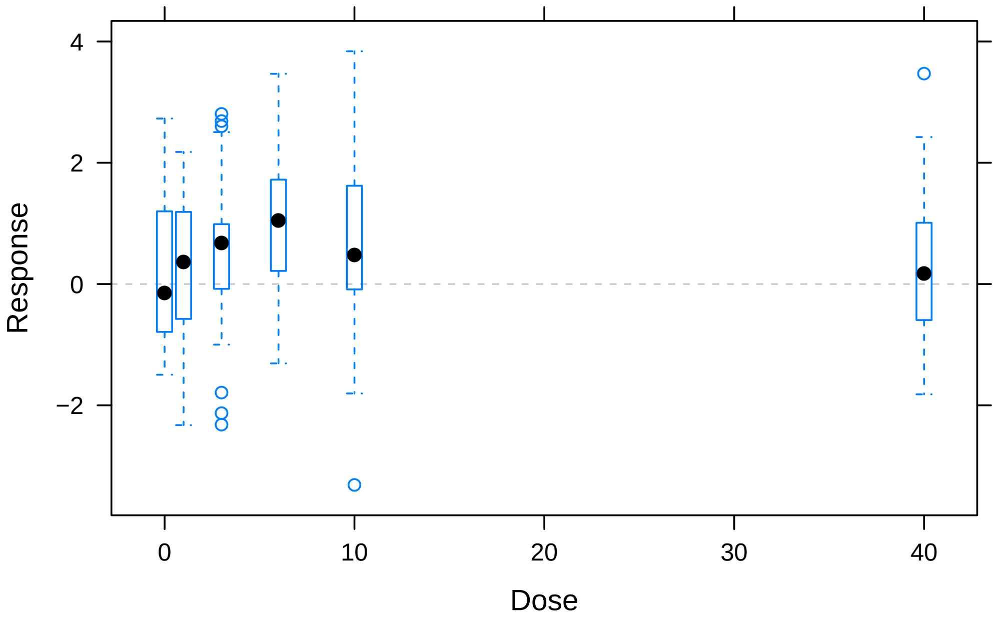
By an initial scan of the Figure 4.1, there seem to be some apparent dose–response relationship in the lower dosage range up to 100 dose units but the dose–response function over the broader range is not very clear.

Table 4.1: Summary of Simulated Study 1

Dose	Sample Size	Mean	Standard Deviation
0	30	0.19	1.220
1	45	0.20	1.257
3	50	0.57	1.161
6	42	0.98	1.159
10	45	0.73	1.393
40	40	0.23	1.208

It is worth to mention that in this simulated study, choosing dose level of 40 being far from the other dosages has its practical meaning. As the cost of developing a new medicine is getting astronomically higher, exploring doses in every potential level is almost infeasible. Spreading a few dosages in a larger spectrum on the scale to improve the odds of finding any dose–response signals

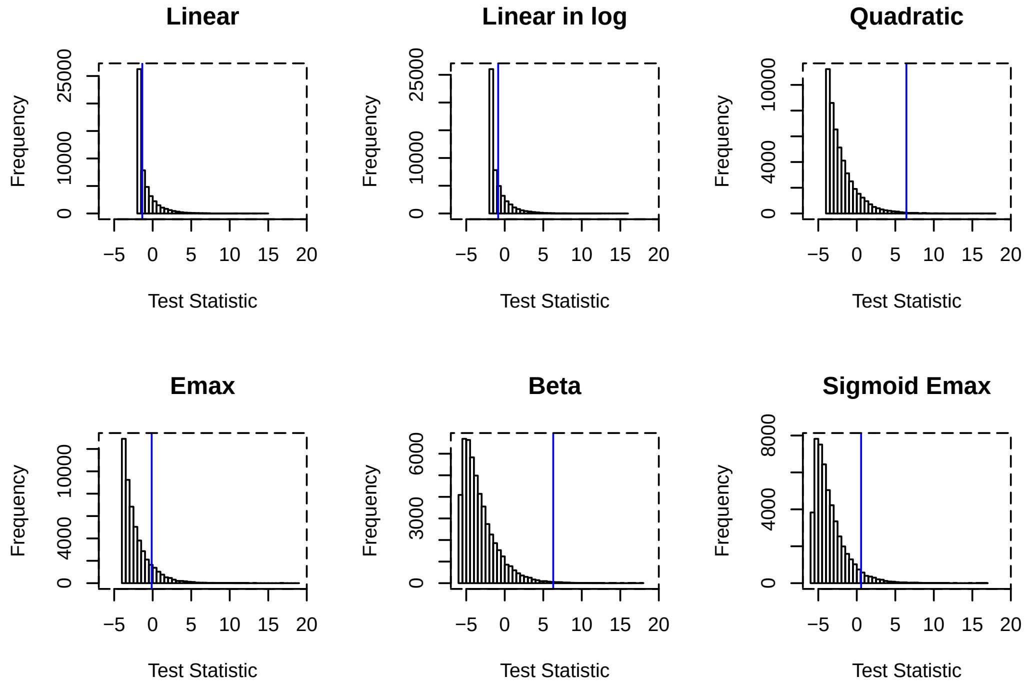
Figure 4.1: Simulated Dose–Response Study 1



is a common approach. It is often to this extent, the “guesstimates” would become less uncertain in specifying the parameters of candidate models.

We treat the simulated study data as our original observations and started the permutation test procedure by selecting a candidate set of six ($M = 6$), commonly used dose–response models: *linear, linear-in-log, quadratic*, $E_{\max}$, *beta* and *sigmoid* $E_{\max}$. We fitted these models to the “original observations” and obtained the observed $\text{LRT}_\ell^{(obs)}$, $l = 1, 2, \ldots, 6$. In assessing the nominal significance of these test statistics, i.e., calculating the corresponding observed $p_\ell^{(obs)}$, we constructed the permutation distributions by fitting the 6 candidate models to $B = 50,000$ permutation samples drawn from the original observations and calculated $\text{LRT}_\ell^{(b)}$ using (4.5) for each permutation sample by each candidate model. Figure 4.2 displays the six corresponding permutation distributions of the LRTs. The observed $\text{LRT}_\ell^{(obs)}$ test statistic value of each model is indicated as the axis mark on which the vertical reference line is plotted. The proportion of the area to the right of the vertical line over the whole histogram area is actually the observed p-values, $p_\ell^{(obs)}$ (4.6).

Figure 4.2: Observed p-value in Simulated Study 1



The relevant statistics from the permutation test procedure are summarized in Table 4.2. They are listed by the order of the observed p-values (from the smallest to the largest). The observed p-values shown in the table nominally indicate that the two models, *Beta* and *Quadratic*, significantly express the possible underlying dose–response signals by the one sided α level of 0.05 and the *Sigmoid* $E_{\max}$ is a borderline significant model. As we discussed before, these marginal (observed) p-values need to be adjusted with respect to the multiplicity. In order to control the overall type I error rate, we performed the resampling-based, single-step multiplicity adjustment procedure described in Section 3.3.1 using (4.7)) and (4.8). We carried out this procedure by calculating the p-value for bth permutation using (4.7) for each model and chose the minimum p-value (4.8) over all model p-values from the same permutation sample. We repeated this calculation for $B = 50,000$ times, thus the permutation distribution of the minimum p-value P_j (3.4) was constructed. If the one sided $\alpha = 0.05$ is the chosen significance level, the single-step adjusted p-values shown in Table 4.2 confirm that the *Beta* and *Quadratic* are the significant models whereas *Sigmoid* $E_{\max}$ (the single-step adjusted $p = 0.1183$) is not a model that should be selected to be in further evaluation for the target dose (MED) estimation.

Table 4.2: Test Statistics of Simulated Study 1

Candidate Model	f_ℓ	$\text{LRT}_\ell^{(obs)}$	Observed $p_\ell^{(obs)}$	Single-step Adjusted $\tilde{p}_\ell$	Step-down Adjusted $\tilde{p}_\ell^*$
Beta	f_5	6.2969	0.0049	0.0136	0.0136
Quadratic	f_3	6.4401	0.0054	0.0150	0.0141
Sigmoid $E_{\max}$	f_6	0.5616	0.0507	0.1183	0.0983
$E_{\max}$	f_4	−0.1324	0.1191	0.2471	0.1970
Linear in log	f_2	−0.8520	0.2887	0.5114	0.3922
Linear	f_1	−1.3510	0.4184	0.6739	0.3922

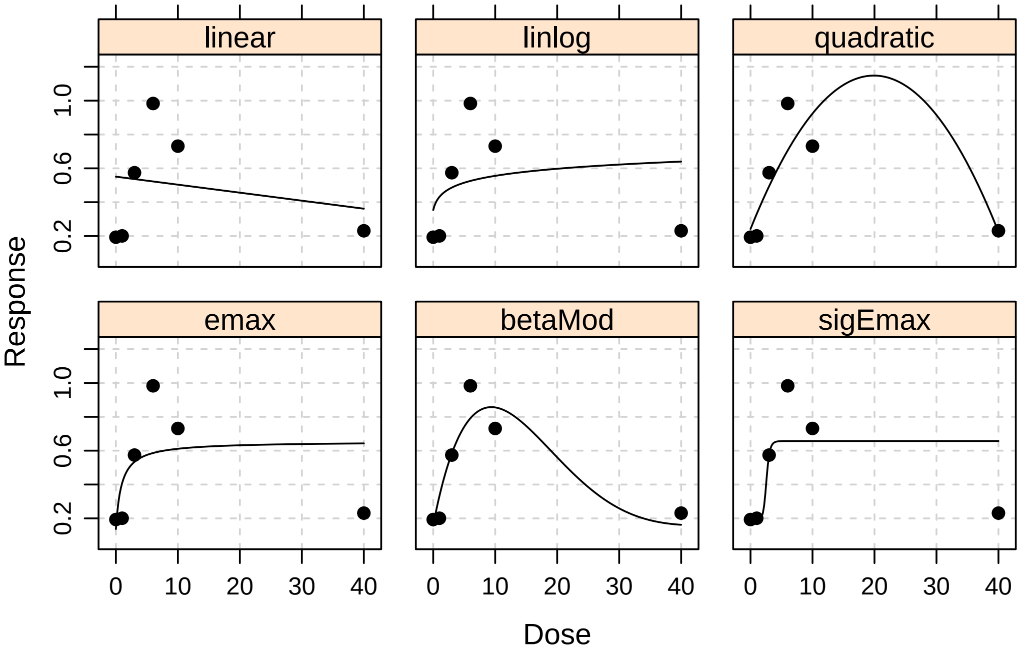
We further explored the more powerful step-down procedure described in Section 3.3.2. The result from the step-down procedure did not change the model selection decision (refer to the adjusted p-values in the last column of Table 4.2). In practice, one would only need to choose one of the adjustment procedures based on research circumstances.

Figure 4.3 displays the six fitted curves to the simulated observations. Visually, the models of *Beta* and *Quadratic* are fitted better as the adjusted p-values suggested.

4.3.2 Data with Weaker Dose–Response

In order to see the performance of our proposed permutation test when dealing with the study observations with weaker dose–response relationships, we investigated the method in a simulated dataset with uncertain dose–response signal. The descriptive summaries of the simulated study are shown in Table 4.3

Figure 4.3: Fitted Dose–Response in Simulated Study 1



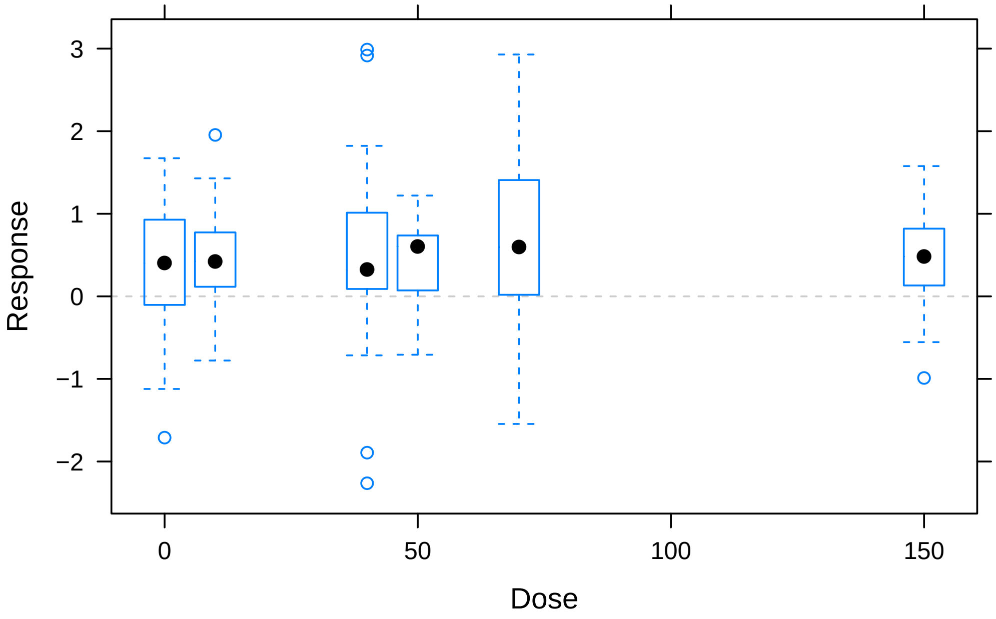
and Figure 4.4. The means of each dosage group seem to suggest that there might be an uptrend dose–response. However, we also observe that the standard deviations of these dose–responses are relatively large in general and vary significantly from group to group. These variations could bring the uncertainty of existences of any underlying dose–response signals.

Table 4.3: Summary of Simulated Study 2

Dose	Sample Size	Mean	Standard Deviation
0	30	0.34	0.769
10	30	0.45	0.654
40	30	0.47	1.125
50	30	0.46	0.470
70	30	0.63	1.026
150	30	0.50	0.604

We used the same candidate set of six models as in the simulated study 1 and performed the test procedure using 50,000 permutations. The observed

Figure 4.4: Simulated Dose–Response Study 2



p-values are plotted in Figure 4.5. The vertical lines on these distributions indicate the highly insignificant nominal p-values.

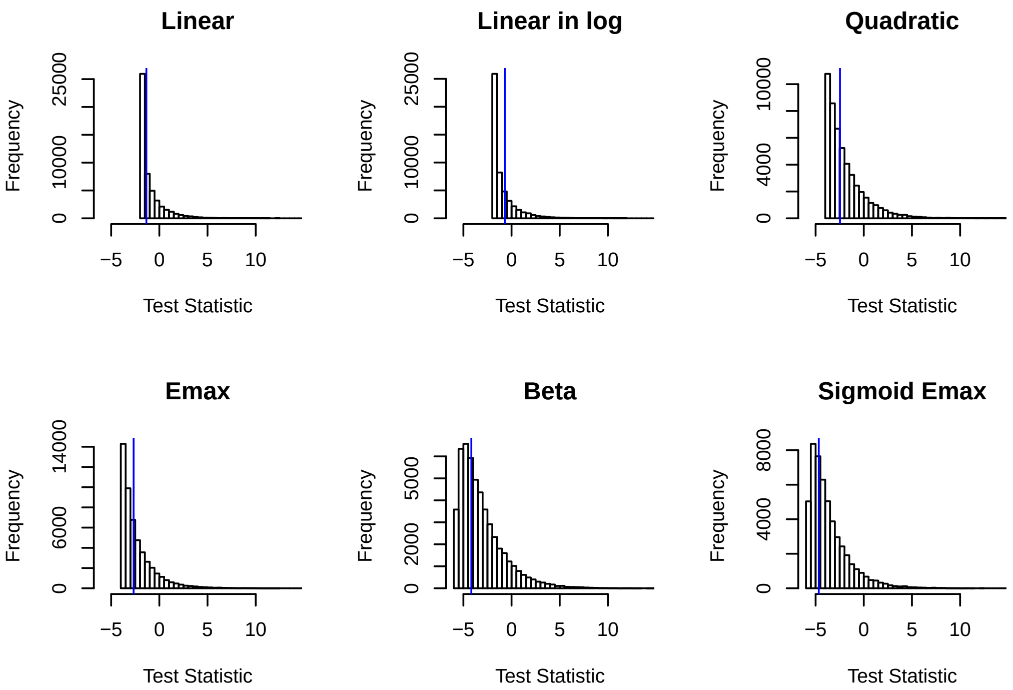
The statistical results shown in Table 4.4 confirm what are illustrated in (Figure 4.6), i.e., there are no obvious dose–response relationships that can be expressed in these model functions.

Table 4.4: Test Statistics of Simulated Study 2

Candidate			Observed	Single-step Adjusted
Model	f_ℓ	$\text{LRT}_\ell^{(obs)}$	$p_\ell^{(obs)}$	$\tilde{p}_\ell$
Linear in log	f_2	-0.7016	0.2564	0.4625
Linear	f_1	-1.3449	0.4211	0.6709
Quadratic	f_3	-2.4728	0.4729	0.7245
$E_{\max}$	f_4	-2.6742	0.4217	0.6716
Beta	f_5	-4.1735	0.5908	0.8333
Sigmoid $E_{\max}$	f_6	-4.6725	0.6294	0.8646

In practice, there may well be cases that dose-dependent effects trend

Figure 4.5: Observed p-value in Simulated Study 2

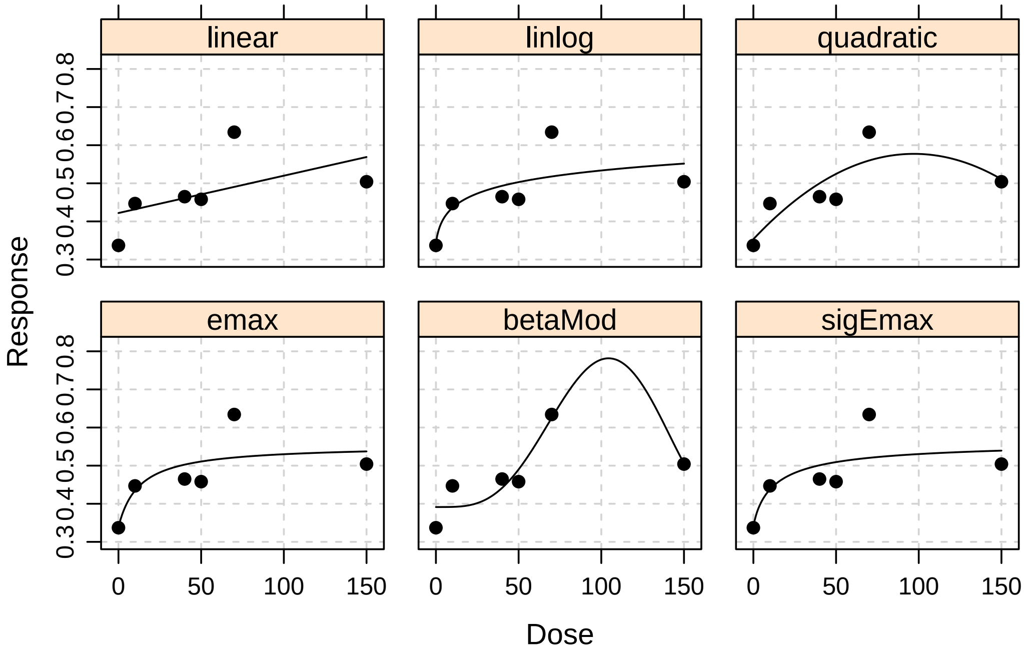


can not be measured quantitatively through statistical models therefore the conventional MCP procedures are more of effective ways to find MED or target doses, if any.

4.3.3 A Dose-Ranging Study for Treating Parkinson's Disease

Neupro® (Rotigotine Transdermal System) is a FDA-approved once-daily prescription patch that is used to treat the signs and symptoms of idiopathic Parkinsons disease (PD) and the moderate-to-severe primary Restless Legs Syndrome (RLS). During its development, a multicenter, randomized, double-blind, placebo-controlled, 5-arm, parallel-group trial was conducted to assess the dose–response of Rotigotine Transdermal System in subjects with advanced-stage PD. The trial was conducted in 2007 and the basic summary

Figure 4.6: Fitted Dose–Response in Simulated Study 2



results are available in ClinicalTrials.gov, a public database maintained by the U.S. National Institutes of Health (NIH). The study identifier in ClinicalTrials.gov is NCT00522379.

Since only the summary information at dosing group level is available, we simulated a full sample of dose and response information at the subject level by using the published efficacy summary statistics from the trial. The simulation was done under the assumption of normal distribution for the efficacy measure in each dosing group. Table 4.5 displays the results from generated sample for the 4 dosing groups and the placebo. The efficacy measure was the change in the absolute time spent “Off” from baseline to the end of 16 weeks of the treatment period as recorded by subject in a daily diary. The Off time means the time-interval that the drug did not seem to be working.

We evaluated a candidate set of eight models: *linear, linear-in-log, quadratic, exponential*, $E_{\max}$, *sigmoid* $E_{\max}$, *beta*, and *logistic* using the proposed permutation test procedure. The results of the test statistics and p-values are listed

Table 4.5: Summary of Neupro® Dose-Ranging Study

Dose (mg/24 hr)	Sample Size	Mean Change in “off” Time (hrs)	Standard Deviation
0	105	−1.50	3.613
2	99	−1.88	2.269
4	103	−2.24	3.297
6	101	−2.29	2.795
8	94	−2.41	2.707

in Table 4.6. The observed p-values suggested all but the *beta* model were significant. After the single-step adjustment was performed, only two models, the *linear-in-log* and *linear*, stood out with $E_{\max}$ model is at the borderline of significance.

Table 4.6: Test Statistics of Neupro® Dose-Ranging Study

Candidate Model	f_ℓ	$\text{LRT}_\ell^{(obs)}$	Observed $p_\ell^{(obs)}$	Single-step Adjusted $\tilde{p}_\ell$	Step-down Adjusted $\tilde{p}_\ell^*$
Linear-in-log	f_2	3.8516	0.0167	0.0426	0.0426
Linear	f_1	3.7364	0.0172	0.0436	0.0426
$E_{\max}$	f_5	2.2814	0.0225	0.0561	0.0483
Exponential	f_4	1.4452	0.0354	0.0846	0.0653
Quadratic	f_3	2.2922	0.0453	0.1070	0.0702
Sigmoid $E_{\max}$	f_6	0.3325	0.0456	0.1076	0.0702
Logistic	f_8	0.3194	0.0483	0.1125	0.0702
Beta	f_7	0.2746	0.0756	0.1680	0.0702

We conducted the more powerful step-down procedure to adjust the observed p-values also. As a result, the $E_{\max}$ model showed its significance and should also be the model in addition to the *linear-in-log* and the *linear* to be evaluated for the target dose estimation.

4.3.4 A Real Dose-Ranging Study with Patient level Data

A double-blind and an active-controlled dose-randing study was conducted on subjects with type 2 diabetes to investigate the dose–response relationship through the clinical efficacy. The efficacy and dosage data was obtained at patient level with no other information being revealed in this dissertation.

The clinical efficacy endpoint was Hemoglobin A1C or HbA1C. A1C is a lab test in which the test measure reflects the mean glucose concentration over the previous period of approximately 8-12 weeks, depending on the individual. The A1C provides a much better indication of long-term glycemic control than the blood and the urinary glucose determinations. Recommendations from the American Diabetes Association (ADA) include the use of HbA1c to diagnose diabetes, using a cut-point of 6.5%. Those who have the AIC above 6.5% are considered diabetic patients.

All subjects included in this study had the A1C level greater than 7% at the beginning of the treatment (at baseline). The goal is to see what the A1C levels would reduce to in different dosage groups compared to a control (an approved drug on the market) at the end of 12 weeks of treatment with daily dose regimen. The summary of doses and the A1C at Week 12 of treatment are displayed in Table 4.7.

Table 4.7: Summary of a Type 2 Diabetes Drug Dosing Study

Dose (mg/24 hr)	Sample Size	HbA1C (%)	Standard Deviation (%)
0	60	6.86	0.926
50	67	7.24	0.840
100	59	7.03	0.843
200	56	6.77	0.625
300	56	6.74	0.786
600	58	6.77	0.737

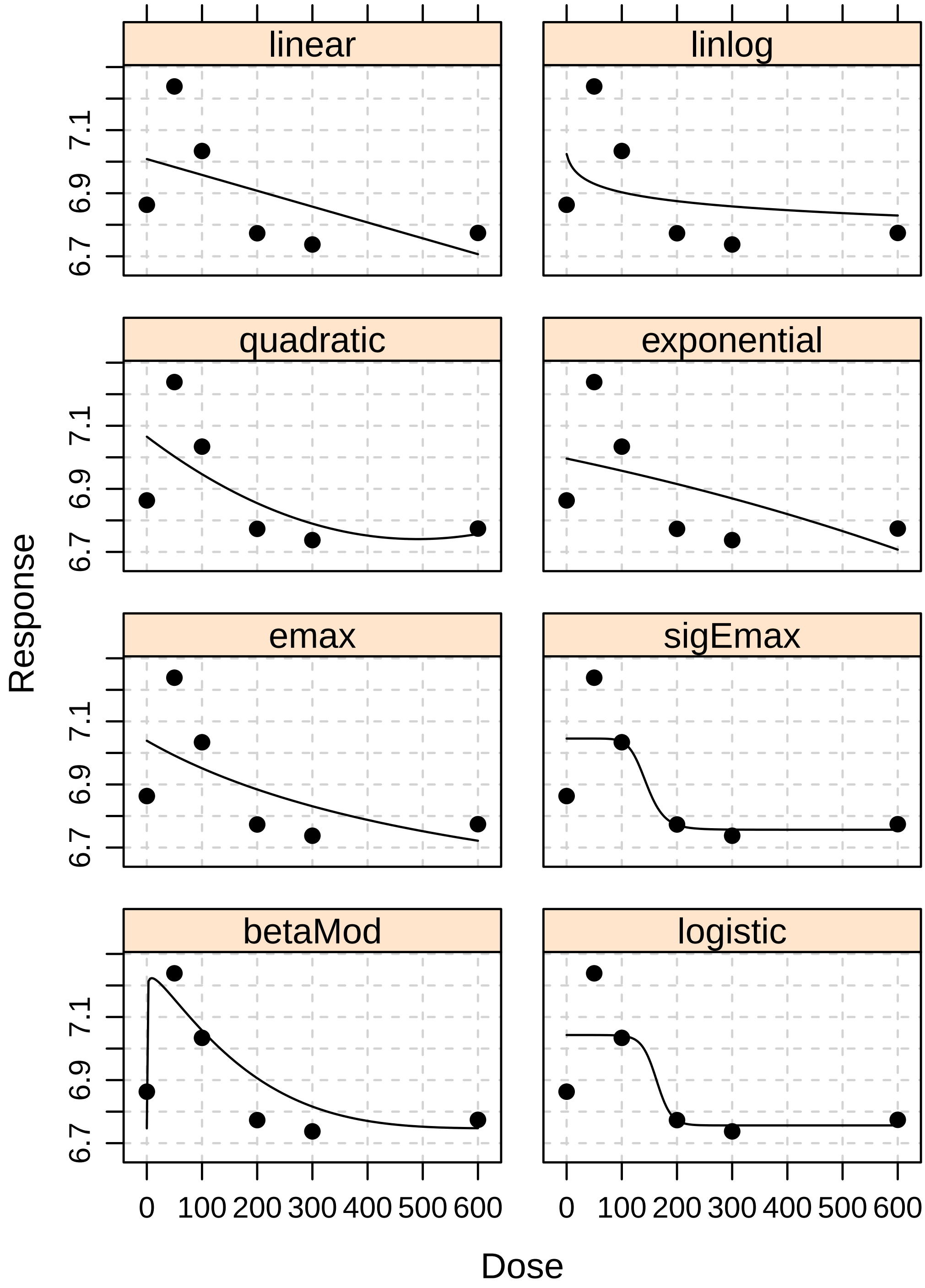
We performed the same permutation test as for the previous study examples, i.e., with the candidate set of 8 models and 50,000 simulations. The statistical analysis results were tabulated in Table 4.8. The marginal p-values were all significant except for the *Linear-in-log* model. If the single-step adjustment procedure is applied, the models of *Beta, Logistic, Sigmoid* and $E_{\max}$ would be chosen with *Linear* model hanging on the borderline when compared to the α level of 0.05. The implementation of the step-down procedure would certainly put the *Linear* model into the consideration of the target dose estimation.

Figure 4.7 display these models fitted to the study observations. Graphically, the Beta model seems to be the best fit in consistent with the adjusted p-values. Note that on the graph, the *Linear* model does not show closely being fitted to the observations but the p-value might have reflected that the penalized LRT is in favor of simpler model.

Table 4.8: Test Statistics of a Type 2 Diabetes Drug Dosing Study

				Single-step	Step-down
Candidate			Observed	Adjusted	Adjusted
Model	f_ℓ	$\text{LRT}_\ell^{(obs)}$	$p_\ell^{(obs)}$	$\tilde{p}_\ell$	$\tilde{p}_\ell^*$
Beta	f_7	6.8814	0.0041	0.0124	0.0124
Logistic	f_8	4.4541	0.0073	0.0209	0.0192
Sigmoid $E_{\max}$	f_6	4.4593	0.0074	0.0214	0.0193
Linear	f_1	3.4314	0.0198	0.0525	0.0459
$E_{\max}$	f_5	1.9716	0.0321	0.0815	0.0679
Exponential	f_4	1.0259	0.0374	0.0940	0.0729
Quadratic	f_3	2.6435	0.0384	0.0963	0.0729
Linear-in-log	f_2	0.1202	0.1489	0.3097	0.0729

Figure 4.7: Fitted Dose–Response in Type 2 Diabetes Study



4.4 Simulation of Power Analysis

In this section we compare the statistical power of detecting significant dose–response signals (models) over a set of candidate models between the MCP–Mod framework (for normally distributed dose–response data) and the RMCP–Mod. As previously discussed, the key difference between the two methods is that in the MCP–Mod, the parameters of certain models in the candidate set need to be pre-specified using "guesstimates" for the optimal contrasts calculation whereas such a specification step is omitted in the RMCP–Mod. With respect to controlling FWER, the MPC–Mod uses the $T_{\max}$ test statistic that follows the multivariate t distribution while the RMCP–Mod employs the LRT through which the permutation distribution (Monte Carlo) of the minimum P-value is constructed. The respective single-stage p-value adjustment procedure is used as the comparing base here.

The simulation study design in the Bretz's paper (2005) [11] was brought over as a reference to guide the power analyses in this section. Five dosage levels ($d = 0,\ 0.05,\ 0.2,\ 0.6,\ 1$), with a single endpoint as the dose–response measured per patient, $Y \sim$ iid. $N(\mu,\ \sigma^2)$ were investigated. Sample sizes per group were set to $n = 15,\ 30,\ 45,\ 60,\ 75$ and 90. The range of smaller sizes (≤ 45) reflects the common practice of conducting phase II clinical trials. The one-sided significance level for PoC (prove that the significant dose–response signal exists) was set at $\alpha = 0.05$. Eight different dose–response shapes for the mean responses $\mu(d)$ were evaluated in the simulations as described in Table 4.9. All of these shapes had properties that at $d = 0$, the standardized response value was about 0.2, and, with an exception of the flat (constant) shape, all had a maximum efficacy response of about 0.8 within the dose interval [0, 1] (i.e., the maximum treatment difference was about 0.6).

The candidate set of the models included commonly used dose–response shapes of *Linear, Linear-in-log, Quadratic, Exponential*, $E_{\max}$ and *Logistic.* The response standard deviation (SD) was set at $\sigma = 1.1$, which, for a group sample size being about $n = 45$, would give a power of 80% for a pairwise test between any two dose levels at the maximum effect of $\delta = 0.6$.

Table 4.9: Specifications of Dose–response Shapes for Simulation

Dose–Response Shape	Parameter Specification
Flat (constant)	0.2
Linear	$0.2 + 0.6d$
Linear in log dose	$0.2 + 0.6\log(5d + 1)/\log(6)$
Quadratic	$0.2 + 2.049d - 1.749d^2$
Exponential	$0.183 + 0.017\exp[2d\log(6)]$
$E_{\max}$	$0.2 + 0.7d/(0.2 + d)$
Logistic	$0.193 + 0.607/\{1 + \exp[10\log(3)(0.4 - d)]\}$
Truncated-logistic	$0.2 + 0.682/\{1 + \exp[10(0.8 - d)]\}$

In order to obtain the “true” underlying dose–response signal, each hypothetical study was simulated from one of models in the candidate set as well as the flat model (for the purpose of preserving the FWER) and the truncated logistic model which was considered a misspecification case for the MCP–Mod (Bretz, at all 2005) [11]. These simulated studies with their model shape specifications are listed in Table 4.9. After the dose–response observations were generated from a particular model shape, they were then evaluated through repetitive simulations using the 6 models in the candidate set for both the MCP–Mod and the proposed permutation test. Except for the constant dose–response shape (the flat model), we would expected high successful rates from the proposed permutation test method performed on these “true” dose–response observations (MCP–Mod was proved in the Bretz’s paper) if it would be effective as we have seen in the previous examples.

The power being evaluated here is based on the definition of the probability that a procedure successfully identifies at least one significant model from the candidate set by comparing the one-side of significance level of 0.05 in simulation trials. There are several challenges of this power comparison. First of all, the power is calculated for the success of different procedures which are not based on some mathematical forms. The simulation approach is often used in this situation as the proportion of the successes out of the total simulations is used to estimate the power. Therefor the power is not "true" power and the bias around the estimation needs to be considered [7]. Secondly, in each simulated study, the statistics of the proposed permutation tests are also estimations themselves since they are obtained from the Monte Carlo distributions based on the resampling set, not the complete set of permutations. Therefore, the error between the Monte Carlo and the exact estimations would be introduced. In addition, the combination of the sizes of the simulations ("outer loop") and the permutation resamples ("inner loop") often requires tremendous computing resources in order to limit the bias of the power estimations and errors in estimating statistics from permutation testing procedures. There are literatures [27] [7] that attempted to give the guidances for optimally choosing the sizes of outer and inner loops to minimize the bias and calculating errors (the terms of "outer loop" and "inner loop" were from these literatures).

Following Boos and Zhang's suggestion [7], we chose 5000 simulated trials (outer loop) per each of the 6 sample sizes for each model shape. For the permutation test, 3000 resamples (inner loop) were randomly drawn from each of the 5000 simulated trials. This is to say that for each model shape with a specific sample size for the 5 dosage groups, 5000 × 3000 × 6 (6 candidate models) calculations were performed. The significantly intensive computational processes were very evident in these power evaluations. Table 4.10 gives the results of these calculations.

The results showed that the power variations on different underling dose–response relationships were consistent between the two methods, i.e., the power was lower in detecting the quadratic form and higher in identifying the logistic trend. The results also showed that the statistical powers for the proposed permutation test procedure were slightly lower across board in the range of the smaller sample sizes but improved to the comparable levels for the MCP–Mod when the sample sizes were increased. This seems to indicate that the proposed approach is slightly less efficient than MCP–Mod method. Note that although the literatures mentioned above pointed out that the larger inner loop size would improve the power in minimizing the bias in estimating the "true power", we could not afford more than 3000 of the permutation resamples as the inner loop size to have reasonable computing times. This limitation on the sampling sizes might reduce the power of the proposed permutation test approach.

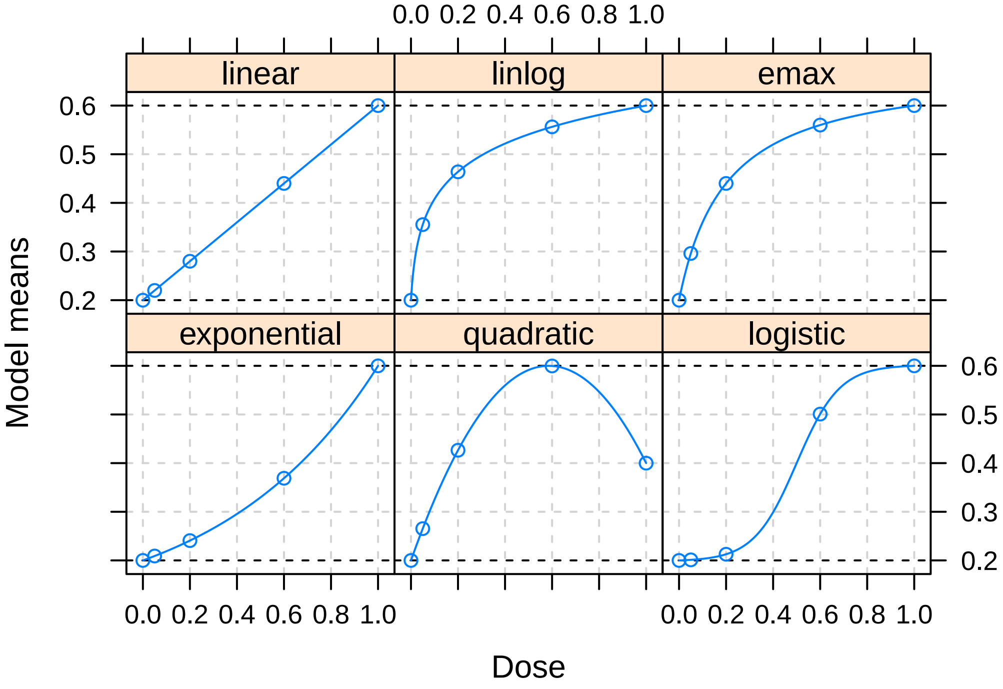
Table 4.10: Results of Power Simulation for PoC

		Simulated data shape							
				Linear in					*Truncated-*
Methods	n	*Flat*	*Linear*	*log Dose*	*Quadratic*	*Exponential*	$E_{\max}$	*Logistic*	*Logistic*
MCP–Mod	15	0.0490	0.4926	0.4988	0.3898	0.4486	0.4762	0.6010	0.4494
	30	0.0564	0.7652	0.7670	0.6492	0.7286	0.7418	0.8676	0.7338
	45	0.0526	0.9022	0.9008	0.8144	0.8838	0.8550	0.9624	0.8910
	60	0.0538	0.9588	0.9610	0.9082	0.9514	0.9542	0.9920	0.9572
	75	0.0514	0.9840	0.9850	0.9560	0.9798	0.9804	0.9970	0.9832
	90	0.0550	0.9952	0.9950	0.9808	0.9934	0.9918	0.9994	0.9942
RMCP-Mod	15	0.0520	0.3542	0.3592	0.2504	0.3306	0.3426	0.4338	0.3444
	30	0.0540	0.6360	0.6376	0.4734	0.6208	0.6166	0.7564	0.6336
	45	0.0498	0.8206	0.8242	0.6608	0.8020	0.8012	0.9112	0.8168
	60	0.0500	0.9230	0.9236	0.7952	0.9128	0.9102	0.9720	0.9238
	75	0.0494	0.9618	0.9642	0.8928	0.9562	0.9598	0.9908	0.9616
	90	0.0538	0.9860	0.9858	0.9364	0.9838	0.9784	0.9982	0.9870

Although both methods can work well for a broad range of underlying dose–response shapes even under some model misspecification situations with the MCP–Mod being more efficient, its use of "guesstimates" may still result in some plausible cases in which the proposed permutation test would have some advantages.

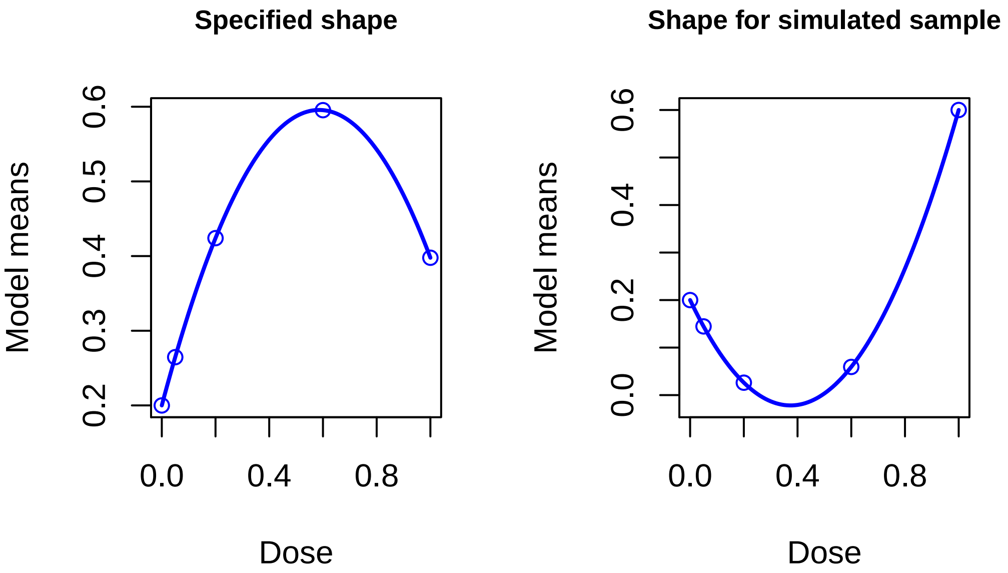
We investigated a simulation study in which the dose–response observations were generated from a convex quadratic form. We formed a set of 6 candidate models (graphically shown in Figure 4.8), i.e., *Linear, Linear-in-log, Quadratic, Exponential*, $E_{\max}$ and *Logistic*, similar to those used in the previous analyses but slightly different in the specifications. Note that in the candidate set, we intentionally have the quadratic form specified as a concave shape whereas the data was generated from a convex dose–response relationship. Figure 4.9 displayed the contrast between the two quadratic forms. The same evaluations with 5000 study simulations from the convex quadratic form for both methods and 3000 resamples in each simulated study for the proposed permutation test were performed. The statistical power for both methods on detecting any dose–response signals (PoC) as well as the probability of the quadratic model being selected were displayed in Table 4.11.

Figure 4.8: Specified Dose–Response Shapes



The power of PoC for the MCP–Mod is consistently lower than the RMCP–Mod except for the simulated study with the smallest sample size. Since the specification for the quadratic model in the MPC-Mod was the concave form, the probability of the quadratic model being significant was consistently small. This indicated that if the PoC was established by MCP–Mod, some other model forms that closely correlated with the convex shape would be selected, such as the *exponential* and the *logistic* both of which have the convex component. This was not the case for the RMPC–Mod, the probability of the quadratic model being significant were in agreement with the power of PoC , i.e., the correct model was selected with higher probability. Therefore, in

Figure 4.9: Quadratic Shapes



situation as such, the RMCP-Mod seems more robust.

Table 4.11: Quadratic Misspecification

	MCP–Mod: 5000 Simulations		Permutation Test (5000 by 3000)	
	Power of	Prob. of Significant	Power of	Prob. of Significant
n	PoC	Quadratic Model	PoC	Quadratic Model
15	0.2446	0.0118	0.2354	0.1630
30	0.4294	0.0084	0.4606	0.3826
45	0.5914	0.0072	0.6492	0.5778
60	0.7212	0.0050	0.7872	0.7256
75	0.8104	0.0066	0.8696	0.8306
90	0.8754	0.0044	0.9348	0.9106

CHAPTER 5

EXTENSION TO BROADER DOSE–RESPONSE TYPES

The extension of the well-developed MCP–Mod methodology to the dose–response signals in binary, counts or time-to-event endpoints as well as repeated measures was developed in the context of general parametric models (Pinheiro et al., 2013) [29]. This extension is named “generalized MCP–Mod”, in short, gMCP–Mod. The key contribution of the gMCP–Mod is to separate the step of modeling dose–response relationships from the step of estimating the corresponding mean dose–response parameters. In another word, the mean dose–responses are estimated based on the relevant parameters from their probability distribution before the dose–response functions being modeled. The reason for this separation approach is that directly maximizing the likelihood (ML) for estimating the model parameters requires derivations of the likelihood in every specific case and would take considerable amount of model-specific coding involvement resulting in practically and computationally inefficient. The theory and detailed implementation of the gMCP–Mod was reviewed in Section 2.3.2.

Without involving with calculations for the optimal contrasts using “guesstimates”, our proposed RMCP-Mod can be easily adapted to the two-stage gMCP–Mod (gRMCP-Mod) for the dose–response signals beyond normally distributed data type. Specifically in the first stage, a generalized linear model for a particular dose–response data type is fitted to the observations so that the estimations of mean dose–response parameters and related variances and covariances can be obtained; in the second stage, the candidate dose–response models are fitted to the mean and variance estimates from the first stage. Trough the second stage of dosing model fitting, the proposed test statistic can be calculated. This two-stage process applies to the original observations as well as to each permutation resample resulting in the permutation distributions being constructed therefore the relevant hypotheses tests can be subsequently performed.

5.1 Choice of Test Statistic

When being applied to the dose–response data other than normally distributed type, the AIC based test statistic, LRT (4.4), in RMCP–Mod needs some modification since direct ML estimation is not used in the two-stage model fitting. One approach is to construct the test statistic using a function of the generalized AIC (gAIC) performed on the candidate models and the flat (constant) model. The gAIC considered as an approach to the model selection criterion was proposed when the gMCP–Mod was developed (Pinheiro et al, 2013) [29]. The gAIC is defined as:

$$\hat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) + \dim(\boldsymbol{\theta})\tau, \text{ where } \hat{\boldsymbol{\Psi}} = (\hat{\boldsymbol{\mu}} - f(\mathbf{x}, \boldsymbol{\theta}))'\hat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{\mu}} - f(\mathbf{x}, \boldsymbol{\theta})).$$

The $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{S}}$ are the estimations of parameters in the dose–response probability distributions from the first stage through the *glm* fitting. The gAIC can be calculated when the dose–response model parameters $\boldsymbol{\theta}$ are estimated in the second stage of the dose–response model ($f(\mathbf{x}, \boldsymbol{\theta})$) fitting process.

Let gAIC$_\ell$ ($\ell = 1, 2, \cdots, M$) denotes the generalized AIC for a candidate model:

$$\begin{aligned} \text{gAIC}_\ell &= \hat{\boldsymbol{\Psi}}(\hat{\boldsymbol{\theta}}) + \dim(\boldsymbol{\theta})\tau \\ &= (\hat{\boldsymbol{\mu}} - f_\ell(\mathbf{x}, \hat{\boldsymbol{\theta}}))'\hat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{\mu}} - f_\ell(\mathbf{x}, \hat{\boldsymbol{\theta}})) + \tau \dim(\boldsymbol{\theta}). \end{aligned} \tag{5.1}$$

In practice, $\tau = 2$ is used to calculate the gAIC value. Refer to the article (Pinheiro et al, 2013) [29] for detailed discussions on gAIC. This gAIC has been implemented as a function in the R package, **DoseFinding**[9].

Similarly, we need to derive a corresponding gAIC for the flat (constant) model for the basis of the model comparisons. Let

$$\begin{aligned} &(\hat{\boldsymbol{\mu}} - f_0(\mathbf{x}, \hat{\boldsymbol{\theta}}))'\hat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{\mu}} - f_0(\mathbf{x}, \hat{\boldsymbol{\theta}})) \\ &\quad = (\hat{\boldsymbol{\mu}} - C \cdot \mathbf{1})'\hat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{\mu}} - C \cdot \mathbf{1}) \\ &\quad = \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{\mu}} - 2C \cdot \mathbf{1}'\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{\mu}} + C^2 \cdot \mathbf{1}'\hat{\boldsymbol{S}}^{-1}\mathbf{1} \\ &\quad = 0, \end{aligned}$$

resulting $\hat{C} = \frac{\mathbf{1}'\hat{\boldsymbol{S}}^{-1}\hat{\boldsymbol{\mu}}}{\mathbf{1}'\hat{\boldsymbol{S}}^{-1}\mathbf{1}}$. and this gives the corresponding AIC calculation for the flat model after the first stage model fitting:

$$\text{gAIC}_0 = (\hat{\boldsymbol{\mu}} - \hat{C} \cdot \mathbf{1})'\hat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{\mu}} - \hat{C} \cdot \mathbf{1}) + k, \tag{5.2}$$

where in common practice, $k = 2$.

From the resampling method perspective, it is more flexible for one to choose test statistics without having the exact probability distribution forms developed since the construction of such distributions can be derived by the permutation resamples. We consider to chose a test statistic for the proposed model selection approach in a form of difference in the gAICs between the flat model and a candidate model (denoted with the subscript D as T_{D_ℓ}) or the ratio of the gAIC from flat model to the gAIC from a candidate model (denoted with the subscript R as T_{R_ℓ}):

$$T_{\text{D}_\ell} = \text{gAIC}_0 - \text{gAIC}_\ell \tag{5.3}$$

or

$$T_{\text{R}_\ell} = \frac{\text{gAIC}_\ell}{\text{gAIC}_0} \tag{5.4}$$

The smaller value close to zero of T_{D_ℓ} would confirm the fact that a null hypothesis is true whereas the larger values would indicate a possible significant dose–response signal. For T_{R_ℓ}, the value would be around one if a null hypothesis is true whereas the lower ratio (< 1) would be a sign of better data fitting for a candidate model. Note that the p-value equations, (4.6) and (4.7) defined in Section 4.1 need to be modified correspondingly by changing the operator sign from $\geq$ to the $\leq$ as:

$$p_\ell^{(obs)} = (1/B)\sum_{b=1}^{B} I(T_{\mathrm{R}_\ell}^{(b)} \leq T_{\mathrm{R}_\ell}^{(obs)}), \tag{5.5}$$

and

$$p_m^{(b)} = (1/B)\sum_{i=1}^{B} I(T_{\mathrm{R}_\ell}^{(i)} \leq T_{\mathrm{R}_\ell}^{(b)}). \tag{5.6}$$

In the follow section, we evaluate the performance of the proposed permutation test approach, gRMCP-Mod against the gMCP–Mod method using these two test statistics. The evaluations are performed through simulation studies with dose–responses in the types of binary and counts data.

5.2 Simulations

The simulated studies in this section are designed to have 6 active dosages of 0, 0.05, 0.2, 0.6, 0.8, 1 with 0 dosage being a control. They are parallel groups with the same sample size in each simulated study. The sampling sizes, ranging from small to large, in different simulated studies are selected for a mimic of realistic Phase II dose–response studies as well as power trend evaluation. The effect size is also considered in choosing sample sizes. In terms of constructing the candidate set of models, we use the following four, commonly used dose–response models (fewer candidate models are selected in the simulation due to the computation resources concern):

We simulate the dose–response observations from one of above four models each time as if it is the underlying true dose–response and use the candidate

Table 5.1: Simulated Dose–Response Shape

Dose–Response Name	Functional Shape
Linear	$f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x$
E_{max}	$f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x/(\theta_2 + x)$
Quadratic	$f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2$
Exponential	$f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1(\exp(x/\theta_2) - 1)$

set of the models (including the simulated model itself) to evaluate the power of detecting the dose–response signals comparing our proposed method to the gMCP–Mod. The flat model is included in the simulation to ensure the type I error for both methods.

5.2.1 Binary Data Type

The binary dose–response data was simulated through a binomial model with the logit scale of $\frac{e^{\eta_i}}{1+e^{\eta_i}}$. The pre-specifications were of $\theta_2 = 0.05$ for the E_{max} model, $\theta_2/\theta_1 = -0.66$ for the *quadratic* model and $\theta_2 = 0.2$ for the *exponential* model. The remaining parameters θ_0 and θ_1 were chosen such that the power for testing the dose with the maximum treatment difference of 0.3 against the control was 80%. With a one-sided significant level at the 5%, the sample size was roughly 30 patients per group. This ensures a realistic range of sample sizes (in terms of the signal to noise ratio) is investigated in the simulations. For each of above 4 models plus the flat dose–response shape, 5000 simulations × 3000 permutations at each sample size of 20, 30, 40 50, 60 were generated for the power evaluation. The list of these models with specific parameter information is displayed in Table 5.2 and the results of the power calculation are presented in Table 5.3.

In the simulation studies with lower range of the samples sizes ($n \leq 40$), the powers for both methods are noticeably varying from one underlying model shape to another with lower power on the E_{max} dose–response data and higher power on the *quadratic* data type. As expected for both methods, the powers

Table 5.2: Binary Dose–Response Shape

Dose–Response Name	η_i
Flat	$\eta_i = -1.734$
Linear	$\eta_i = -1.734 + 1.5334 \times d_i$
Quadratic	$\eta_i = -1.734 + 4.0842 \times d_i - 2.7094 \times d_i^2$
Exponential	$\eta_i = -1.734 + 0.0104 \times [\exp(d_i/0.2) - 1]$
$E_{\max}$	$\eta_i = -1.734 + 1.61 \times d_i/(0.05 + d_i)$

Table 5.3: Binary Response Simulation

	Simulated data shape					
Methods	n	*Flat*	*Linear*	*Quadratic*	*Exponential*	E_{max}
gMCP–Mod	20	0.0218	0.7848	0.8218	0.7292	0.6862
	30	0.0310	0.9310	0.9614	0.8894	0.8776
	40	0.0318	0.9802	0.9896	0.9624	0.9626
	50	0.0432	0.9948	0.9982	0.9886	0.9884
	60	0.0488	0.9974	0.9994	0.9932	0.9974
gRMCP–Mod	20	0.0462	0.7364	0.7454	0.6996	0.5676
Test (T_{D_ℓ})	30	0.0486	0.8912	0.9146	0.8536	0.7934
	40	0.0452	0.9640	0.9752	0.9372	0.9156
	50	0.0486	0.9888	0.9942	0.9798	0.9680
	60	0.0488	0.9958	0.9984	0.9896	0.9900
gRMCP–Mod	20	0.0468	0.7318	0.7340	0.6984	0.5558
Test (T_{R_ℓ})	30	0.0480	0.8826	0.9066	0.8496	0.7776
	40	0.0436	0.9582	0.9700	0.9326	0.9056
	50	0.0506	0.9866	0.9932	0.9794	0.9624
	60	0.0520	0.9956	0.9982	0.9892	0.9872

consistently increase as the sample size gets larger. Overall, for the proposed permutation test method using the test statistic, T_{D_m}, the powers are slightly lower than gMCP–Mod particularly in the lower range of the sample sizes. For the test statistic T_{R_m}, the powers are slightly lower for the proposed approach than for the gMCP–Mod.

5.2.2 Counts Data Type

A similar evaluation was performed on the overdispersed counts data using the negative binomial with the probability of $\frac{1}{1+\mu_i}$ where μ_i is listed in Table 5.4.

Table 5.4: Counts Dose–Response Shape

Dose–Response Name	μ_i
Flat	$\mu_i = e^2$
Linear	$\mu_i = e^{2-0.5018\times d_i}$
Quadratic	$\mu_i = e^{2-1.635\times d_i+1.2456\times d_i^2}$
Exponential	$\mu_i = e^{2-0.003466\times[\exp(d_i/0.2)-1]}$
$E_{\max}$	$\mu_i = e^{2-0.5427\times d_i/(0.05+d_i)}$

Only the test statistic T_{D_m} was used as the preliminary testing run did not indicate that the T_{R_m} would render any different simulation results, if not slightly less efficient. 4000 simulations × 3000 permutations at each sample size of 45, 60, 75 90 and 105 are generated as the larger sample sizes are needed for evaluating the statistical power through the dose–response in counts data type.

The results from these simulations are reported in Table 5.5. The lower powers of the RMCP–Mod compared to the gMCP–Mod were more pronounced than those seen in the simulated binary dose–response data type. Besides the bias possibly coming from the limitation to the sizes of the simulations combined with the sizes of the Monte Carlo samples, the test statistic that seems to work well with binary data may not be suitable for the counts data. This points out that further research is need to either modify the proposed test statistic or search for and evaluate new ones.

Table 5.5: Simulation of Dose–Response in Counts

	Simulated data shape					
Methods	n	*Flat*	*Linear*	*Quadratic*	*Exponential*	E_{max}
gMCP–Mod	45	0.0330	0.5310	0.5798	0.4893	0.4458
	60	0.0419	0.6738	0.7415	0.6245	0.5828
	75	0.0410	0.7653	0.8243	0.7100	0.6902
	90	0.0423	0.8313	0.8885	0.7870	0.7695
	105	0.0433	0.8838	0.9313	0.8415	0.8415
RMCP–Mod	45	0.0480	0.4548	0.4780	0.4390	0.3445
Test (T_{D_ℓ})	60	0.0453	0.5903	0.6198	0.5558	0.4698
	75	0.0505	0.6828	0.7225	0.6485	0.5745
	90	0.0510	0.7580	0.7985	0.7165	0.6683
	105	0.0500	0.8313	0.8633	0.7803	0.7425

CHAPTER 6

CONCLUSION AND FUTURE WORK

In the previous chapters, we developed an alternative statistical method to the MCP–Mod/gMCP–Mod framework [11] [29] in identifying the significant model(s) that possibly best represent(s) the underling dose-response relationship reflected in the study observations. The proposed RMCP–Mod/gRMCP–Mod formulates the model evaluations into a multiple hypotheses issue and solve it using the resampling based permutation tests and the corresponding multiplicity procedures developed by other researchers [38] [36]. In the meantime, we demonstrated the advantage of this resampling based method in the avoidance of using the "guesstimates", a source of potential model uncertainties introduced in the MCP–Mod/gMCP–Mod.

The RMCP–Mod/gRMCP–Mod was successfully performed on the hypothetical simulated studies and evaluated in reanalyzing the real dose-ranging clinical trials. The results of these performances generally suggested the validity of the proposed method.

We made the attempt to compare the statistical power between the proposed approach and the MCP–Mod on the normally distributed dose-response data although challenges are presented in such power comparisons as the power calculations tend to be based on the simulations compounded by the Monte

Carlo permutation resamples. Similar power comparisons were made to have evaluated both approaches on the binary and counts data. Overall, the proposed method tracks closely to the power of MCP–Mod/gMCP–Mod on the normally distributed and binary type of dose-response data. However, the more efficient results were seen in the MCP–Mod/gMCP–Mod, particularly in the range of small sample sizes of dosage groups. On the counts data type, the gMCP–Mod performed better than the gRMCP–Mod in terms of the statistical power. This may not discount the practical use of the proposed approach since in the stage of detecting dose-response signals, researchers often trade off the type I error rate with the power. Our power comparisons in this dissertation adopted a relatively stringent significance level of α at 0.05. One could increase the power of the proposed procedure even on the counts data by relax the significance level a bit.

The computational limitation may also play the role of the power calculation when the number of permutation resamples was restricted to the relative smalls size (small Monte Carlo sampling size) of a few thousands in each simulation study. There is literature [27] [7] in which authors discussed the bias arising from simulations and resampling procedures and how to minimize them by optimally choosing the number of simulations (outer loop) and the number of permutations (inner loop). The methods discussed in these literatures only provide the guidances, one still needs to evaluate the circumstances for more complicated cases.

Throughout this dissertation research, we observed some limitations to the proposed permutation tests. These limitations prompt the future work that can be explored to improve the method:

- Differentiate highly significant models - When a strong dose-response signal exits in the data, the propose approach, same as the MCP–Mod/gMCP–Mod, would tend to arrive at many highly significant models even with the FWER being controlled. The decision to chose the most significant model for further evaluation on the estimation of the target dose or MED would become difficult. The model averaging may be one of the solutions. However, with the cost of computational resources continuously decreases, computational approach such as combining the permutation test with the cross validation technique may be worthy of exploration for model selection process. If the sample size of a dose ranging study is relatively large, the dataset can be randomly split into two parts. One part can be used for model identifying purpose, the other part can be used for the validation of the identified model. This is again may be done by using a permutation test. The test statistic could be a goodness-of-fit type of metric discussed in Section 12.4 of the book "*Permutation, Parametric, and Bootstrap Tests of Hypothesis*" by Phillip Good [19].

- More efficient test statistic - Although the simulations compounded by the permutation resamples make the power estimation a challenging task, the test statistic may still have room to be improved upon. The evaluations of the characteristics of the permutation distributions for different type of data may worth to exam.

- Evaluations on more data types - The RMCP–Mod should be further evaluated and researched for the performances on types of data that were studied through the gMCP–Mod such as the time-to-event and the over-time repeated measures for more broad robust results.

Overall the proposed permutation test approach, RMCP–Mod/gRMCP–Mod seems to be a viable method. Further research on its robustness to the variety of different dose-response measurements are needed for the completeness of its development.

REFERENCES

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

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and Casaki, editors, *Second international symposium on information theory*, pages 267–281, 1973.

[3] HPC at Temple University. *The Owl's Nest User Guide*. `http://www.hpc.temple.edu/owlsnest/OwlsnestUserGuide.html`.

[4] D.M. Bates and D.G. Watts. *Nonlinear regression analysis and its applications*. Wiley, 1988.

[5] P. Bauer. A note on multiple testing procedures for dose finding. *Biometrics*, (53):1125–1128, 1997.

[6] R. C. Blair and W. Karniski. Distribution-free statistical analyses of surface and volumetric maps. In R. W. Thatcher, M. Hallett, T. Zeffiro, E. R. Jony, and M. Huerta, editors, *Functional Neuroimaging: Technical Foundations*, pages 19–28, 1994.

[7] Dennis D. Boos and Ji Zhang. Monte carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association*, (95 450):486–492, 2000.

[8] B. Bornkamp. Comparison of model-based and model-free approach for the analysis of dose response studies. *Diploma thesis*, 2006.

[9] Bjoern Bornkamp, José Pinheiro, and Frank Bretz. *DoseFinding: Planning and Analyzing Dose Finding experiments*, 2014. `http://cran.r-project.org/web/packages/DoseFinding/index.html`.

[10] F. Bretz, A. Genz, and L.A. Hothorn. On the numerical availability of multiple comparison procedures. *Biometrical Journal*, (43):645–656, 2001.

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

[12] S.T. Buckland, K.P. Burnham, and N.H. Augustin. Model selection: an intergral park of inference. *Biometrics*, (53):603–618, 1997.

[13] J.M. Chambers and T.J. Hastie. *Statistical models in S*. Chapman and Hall, 1992.

[14] A. Dmitrienko, A.C. Tamhane, and F. Bretz. *Multiple testing problems in pharmaceutical statistics*. Taylor and Francis Group, 2010.

[15] Meyer Dwass. On the asymptotic normality of some statistics used in non-parametric tests. *The Annals of Mathematical Statistics*, (26):334–339, 1955.

[16] Dirk Eddelbuettel. *CRAN Task View: High-Performance and Parallel Computing with R*, 2014. `http://cran.r-project.org/web/views/HighPerformanceComputing.html`.

[17] R.J. Freund and R.C. Littell. *SAS System for Regression*. SAS Institute, third edition, 2000.

[18] A. Genz and F. Bretz. *Computation of multivariate normal and t-Ppobabilities*. Springer Verlag, 2009.

[19] Phillip Good. *Permutation, parametric, and bootstrap tests of hypotheses.* Springer, thrid edition, 2005.

[20] Westfall P. H. and J. F. Troendle. Multiple test with minimum assumptions. *Biometrical Journal*, (50 5):745–755, 2008.

[21] Tim Hesterberg. Bootstrap methods and permutation tests. In *The Practice of Business Statistics: Using Data for Decisions*, chapter 18, pages 18–1 to 18–74. W. H. Freeman, 2nd edition, 2003.

[22] T. Hothorn, F. Brezts, and P. Westfall. Simultaneous inference in general parameteric models. *Biometrical Journal*, (50):346–363, 2008.

[23] Karl-Heinz Jöckel. Finite sample properties and asymptotic efficiency of monte carlo tests. *The Annals of Statistics*, (14):336–347, 1986.

[24] B. Klingenberg. Proof of concept and dose estimation with binary responses under model uncertainty. *Statistics in Medicine*, (28):274–292, 2009.

[25] M.R. Law, N.J. Wald, and A.R. Rudnicka. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *BMJ*, (326):1423–1429, 2003.

[26] Norman Matloff. *Parallel Computation in Data Science: With Examples in R.* Chapman and Hall/CRC, 2015.

[27] Neal L. OdenSource. Allocation of effort in monte carlo simulation for power of permutation tests. *Journal of the American Statistical Association*, (86 416):1074–1076, 1991.

[28] J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS.* Springer, 2000.

[29] José Pinheiro, Bjorn Bornkamp, Ekkehard Glimmb, and Frank Bretz. Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine*, 33(10):1646–1661.

[30] W. Schaafsma and L.J.W. Smid. Most stringent somewhere most powerful tests against alternatives restricted by a number of linear inequalities. *The Annals of Mathematical Statistics*, (37):1161–1172, 1963.

[31] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, (6):461–464, 1978.

[32] G.A.F. Seber and C.J. Wild. *Nonlinear regression.* Wiley, 1989.

[33] H. Shimodaria. An application of multiple comparison techniques to model selection. *Ann. Inst. Stat. Math.*, (50):1–13, 1998.

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

[35] Naitee Ting, editor. *Dose finding in drug development.* Springer, 2006.

[36] J. F. Troendle. A stepwise resampling method of multiple hypothesis testing. *JASA*, (90 (429)):370–378, 1995.

[37] Simon Urbanek. *multicore: A stub package to ease transition to 'parallel'*, 2014. `http://cran.r-project.org/web/packages/multicore/index.html`.

[38] Young S. S. Westfall, P. H. *Resampling-based multiple testing: examples and methods for p-value adjustment.* John Wiley, 1993.

[39] W. Zucchini. An introduction to model selection. *Math. Psychol*, (44):41–61, 2000.

APPENDIX A

Computational Resources

All the computational procedures in this dissertation were performed in R with R package, **DoseFinding**, developed by Bjoern Bornkamp, José Pinheiro and Frank Bretz for MCP-Mod/gMCP-Mod framework [9]. The parallel technology has to be used for this dissertation due to the large size of combination of outer and inner loops in sample studies and simulations. The **multicore** package [37] with parallel tool (using multiple CPUs) in R was chosen for these calculations. The detailed theory of parallel computing is discussed in the book, "*Parallel Computing for Data Science: With Examples in R*" [26].

A.1 Computing Times without "multicore" Package

Tremendous computational time was spent in this dissertation. Currently under Windows environment, one can not take advantages of the R package **multicore** since it runs only on Unix-family (e.g. Linux and Mac) platforms. Table A.1 displays the computational times for running an R function named "foo" under the Windows environment; note that the **multicore** parameter can only be set to 1 (mc.cores=1). The function "foo" consists of permutation testing and model fitting procedures that served the main computations for the sample studies and simulations throughout the dissertation. In these test runs, the function executed by taking the outer loops of 10 or 100 (simulation size) with corresponding inner loops of 100 or 1000 (resample size for each simulation) for evaluating 6 candidate models with normally distributed dose-response data. The total process time (elapsed time) was approximately 43 seconds for the simulation size of 10 with 100 resamples per each simulation; and was about 3184 seconds (53 minutes) for the simulation size of 100 with 1000 resamples per each simulation. The simulations would easily require hours and even days to finish if the sizes of outer and inner loops increase.

Table A.1: Simulations with Single Core in R under Windows

```
> library(parallel)

> rslt <- mclapply(1:10, ## the number of simulations
+            FUN=function(x, n, m, doses, shape)
+            foo(n, m, doses, shape),
+            n=30, m=100, doses=doses, shape=5, mc.cores=1)

> proc.time()
   user  system elapsed
  32.54    0.87   43.13

> rslt <- mclapply(1:100, ## the number of simulations
+            FUN=function(x, n, m, doses, shape)
+            foo(n, m, doses, shape),
+            n=30, m=1000, doses=doses, shape=5, mc.cores=1)

> proc.time()
   user  system elapsed
3155.66    1.38 3184.44
```

A.2 Temple University High Performance Computing Cluster

Since the R package **multicore** runs only on Unix-family platforms, not Windows, the computations in this dissertation were mainly performed on the Temples' High Performance Computing (HPC) environment, which was funded in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854. Temple offers high performance computing Linux environments. The Owl's Nest is one of them to provide university researchers sufficient computing power to do their intensive data analysis. For using Owl's nest cluster, the user guide [3] lists the computers, technical specifications and requirements. It provided guidance for completing all the computational work in this dissertation, which was performed on this cluster, under the planned schedules.

It appeared that through running small scaled tests for the function "foo", the computers with about 10 cores (CPUs) on the Temple Owl's Nest would reduce the elapsed time by a factor of ten. For example, we specified the number of cores to 12 (mc.scores=12) for the same simulation as show in Table A.1, being executed on one of the computers on the cluster. The total process time (elapsed) was reduced to only 315 seconds, a little more than 5 minutes, for the simulation size of 100 with 1000 resamples compared to 53 minutes used under the Windows.

Table A.2 illustrates the computing time used on the cluster for one of the dissertation simulations. The simulation was performed for evaluating 6 candidate models on the simulated E_{max} dose-response data with a sample size of 90 for each of the 5 dosage groups in the study. One of the computers on the cluster with 48 cores was available at the scheduled time. The outer loop size was set to 5000 with each loop had 3000 inner loops (resamples). This particular simulation analysis used the elapsed time of 40,853 seconds, about 11 hours 20 minutes. The sum of the user (CPU) and system time suggest that on the same computer without using the R package **multicore**, this simulation would take almost 400 hours, more than 16 days, to finish.

Table A.2: Simulations with Multiple Cores in R under Linux

```
> set.seed(1963)
> nsim=5000
> SEEDs <- round(runif(nsim)*100000)

> ## call the function foo on multicore system
> rslt <- mclapply(1:nsim, ## the number of simulations
+            FUN=function(x, n, m, doses, shape, seed) {
+            set.seed(SEEDs[x])
+            foo(n, m, doses, shape)},
+            n=90, m=3000, doses=doses, shape=4, seed=SEEDs,
mc.cores=48)

> proc.time()
       user      system     elapsed
1401960.167     946.496   40852.967
```