ICSA Book Series in Statistics
Series Editors: Jiahua Chen · Ding-Geng Chen

Naitee Ting
Ding-Geng Chen
Shuyen Ho
Joseph C. Cappelleri

Phase II Clinical Development of New Drugs

Springer

ICSA Book Series in Statistics

More information about this series at http://www.springer.com/series/13402

Naitee Ting · Ding-Geng Chen
Shuyen Ho · Joseph C. Cappelleri

Phase II Clinical Development of New Drugs

Springer

Naitee Ting
Boehringer-Ingelheim Pharmaceuticals Inc.
Ridgefield, CT
USA

Shuyen Ho
PAREXEL International
Durham, NC
USA

Ding-Geng Chen
University of North Carolina
Chapel Hill, NC
USA

and

University of Pretoria
Pretoria
South Africa

Joseph C. Cappelleri
Pfizer, Inc.
Groton, CT
USA

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-981-10-4192-1 ISBN 978-981-10-4194-5 (eBook)
DOI 10.1007/978-981-10-4194-5

Library of Congress Control Number: 2017934053

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Clinical development of new drugs (or new biologics) can be broadly divided into four phases: Phases I, II, III, and IV. The first 3 phases are considered as "pre-marketing development," and Phase IV is known as the "post-marketing" phase. For the pre-marketing clinical development programs, concerns for developing agents treating life-threatening diseases, such as cancer, are different from concerns for other indications. In oncology development programs for treating cancer patients, typical Phase I trials recruit cancer patients and escalate doses to help find the maximally tolerable dose (MTD). Then, this MTD is used in Phase II for proof of concept (PoC) and for Phase III confirmatory trials.

On the other hand, in most of the product candidates developed for the treatment of chronic diseases or non-life-threatening conditions, Phase I clinical trials recruit healthy normal volunteers, and the main objectives of Phase I are to understand the pharmacokinetics (PK) and to estimate MTD. Phase II clinical trials are designed to establish PoC and to estimate the efficacy dose range. After these exploratory phases (Phase I to explore PK and upper limit of dose range, Phase II to explore lower limit of dose range), Phase III uses large-scale confirmatory clinical trials with an objective to establish the longer-term efficacy and safety of this new candidate product. If the dose(s) studied in these Phase III trials demonstrate efficacy and safety, then these doses can be approved by regulatory agency for indicated patient population use. The focus of this book is on the Phase II development of medicinal products treating most non-life-threatening indications, oncology generally excluded (although some of the material may be relevant to oncology trials).

In practice, most of the new compounds or new biologics fail to become a successful medical intervention. Hence, an appropriate angle to view clinical development could be to think of it as a weeding out process. In other words, if a product candidate does not demonstrate clinical efficacy or safety, then this candidate should be stopped for further development. From this point of view, an efficient clinical development process should be such that if the candidate is a successful drug, then this drug should be made available for treating patients as soon as possible. On the contrary, if this candidate is not a successful agent, then it

should be stopped from further development as early as possible, in order to avoid unnecessary sponsor investment and resources.

The difficult question is how to achieve both objectives. We believe the most efficient process of new drug development should align with the concept of "design to stop." Using this strategy, the project team uses every clinical trial to weed out the drug candidate under development. This is the most efficient way of ensuring an unsuccessful candidate to be stopped at the earliest stage. The natural question would then be "What about the successful candidates?" In fact, a good drug reveals its good properties under most simple designs. Across all the successful drugs on the market, most of them revealed their good efficacy and safety characteristics at very early stage of clinical development. In other words, with the "design to stop" strategy, it is very unlikely to weed out a very good product candidate.

For study drugs developed to treat a non-life-threatening condition, before it enters Phase II, it is not known whether this study drug can deliver the expected efficacy (because efficacy cannot be observed from healthy volunteers in Phase I). At the end of Phase II, a critical decision must be made as to whether to continue developing this test drug into Phase III, which implies an enormous amount of commitment in investment and resources. Given this background of medicinal product development, project team members need to know that among all of the product development processes, no other step could be more important and critical than Phase II.

The objective of this book is to clarify the importance of Phase II clinical development and also to lay out the key thinking process in designing and analyzing of each Phase II clinical study. With a clear understanding of the importance of Phase II, as well as how to face these challenges in new product development, readers will be able to design the most efficient Phase II clinical trials. As a product candidate moves into Phase II, these strategies will maximize the likelihood that a "not so good" candidate will be weeded out as early as possible and that the really good product can reach to the patients as soon as possible.

As a general note, the references for each chapter are at the end of the chapter so that the readers can readily refer to the chapter under discussion. Thus, each chapter is self-contained with respect to references. The first seven chapters focus on design of Phase II clinical trials, and then, Chaps. 8–10 discuss analysis of these clinical data. Finally, Chap. 11 covers a Bayesian approach, and Chap. 12 lays out Phase III trial considerations.

The entire content of this book is intended solely and strictly for educational and pedagogical purposes. The material herein expresses the views of the authors and does not in any way reflect the views of their employers or any other entity.

We would like to express our gratitude to many individuals. Thanks to Hannah Qiu (Springer/ICSA Book Series coordinator) from Springer Beijing for their professional support to make this book published at Springer (http://www.springer.com/series/13402).

We welcome any comments and suggestions on typos, errors and future improvements about this book. If such an exchange, please contact Dr. Naitee Ting (e-mail: naitee.ting@boehringer-ingelheim.com) as the corresponding author and,

if desired, Drs. Ding-Geng Chen (e-mail: DrDG.Chen@gmail.com or dinchen@email.unc.edu), Shuyen Ho (e-mail: shuyenho@gmail.com), and Joseph C. Cappelleri (e-mail: joseph.c.cappelleri@pfizer.com).

Ridgefield, CT, USA — Naitee Ting, Ph.D.
Chapel Hill, NC, USA/Pretoria, South Africa — Ding-Geng Chen, Ph.D., M.S.
Durham, NC, USA — Shuyen Ho, Ph.D.
Groton, CT, USA — Joseph C. Cappelleri, Ph.D., M.P.H., M.S.
January 2017

Contents

About the Authors

Naitee Ting is a Fellow of ASA. He is currently a Director in the Department of Biostatistics and Data Sciences at Boehringer-Ingelheim Pharmaceuticals Inc. (BI). He joined BI in September of 2009, and before joining BI, he was at Pfizer Inc. for 22 years (1987–2009). Naitee received his Ph.D. in 1987 from Colorado State University (major in Statistics). He has an M.S. degree from Mississippi State University (1979, Statistics) and a B.S. degree from College of Chinese Culture (1976, Forestry) at Taipei, Taiwan.

Naitee published articles in Technometrics, Statistics in Medicine, Drug Information Journal, Journal of Statistical Planning and Inference, Journal of Biopharmaceutical Statistics, Biometrical Journal, Statistics and Probability Letters, and Journal of Statistical Computation and Simulation. His book "Dose Finding in Drug Development" was published in 2006 by Springer, and is considered as the leading reference in the field of dose response clinical trials. The book "Fundamental Concepts for New Clinical Trialists", co-authored with Scott Evans, was published by CRC in 2015. Naitee is an adjunct professor of Columbia University, University of Connecticut and University of Rhode Island. Naitee has been an active member of both the American Statistical Association (ASA) and the International Chinese Statistical Association (ICSA).

Ding-Geng Chen is a fellow of the American Statistical Association and currently the Wallace Kuralt distinguished professor at the University of North Carolina at Chapel Hill, USA, and an extraordinary professor at University of Pretoria, South Africa. He was a professor at the University of Rochester and the Karl E. Peace endowed eminent scholar chair in biostatistics at Georgia Southern University. He is also a senior consultant for biopharmaceuticals and government agencies with extensive expertise in clinical trial biostatistics and public health statistics. Professor Chen has written more than 150 referred publications and co-authored/co-edited twelve books on clinical trial methodology with R and SAS, meta-analysis using R, advanced statistical causal-inference modeling, Monte-Carlo simulations, advanced public health statistics and statistical models in data science.

Shuyen Ho received his Ph.D. in Statistics from University of Wisconsin—Madison, and his Bachelor in Applied Mathematics from Taiwan. Dr. Ho is a Biostatistics Director at PAREXEL International in Durham, North Carolina and has worked in the pharmaceutical industry for over 25 years. Prior to PAREXEL, he was a Clinical Statistics Director at GlaxoSmithKline (GSK) and Group Leader at Merck. He specializes in Phase II & III clinical development and has helped developed widely used respiratory medicines such as Claritin, Advair and Veramyst.

Joseph C. Cappelleri earned his M.S. in statistics from the City University of New York (Baruch College), Ph.D. in psychometrics from Cornell University, and M.P.H. in epidemiology from Harvard University. Dr. Cappelleri is a senior director of biostatistics at Pfizer Inc. He has also served on the adjunct faculties at Brown University, Tufts Medical Center, and the University of Connecticut. A Fellow of the American Statistical Association, he has delivered numerous conference presentations and published extensively on clinical and methodological

topics, including regression-discontinuity designs, meta-analysis, and health measurement scales. Dr. Cappelleri is the lead author of the book "Patient-Reported Outcomes: Measurement, Implementation and Interpretation."

Chapter 1
Introduction

1.1 Background

Most of the drugs available in pharmacy started out as a chemical compound or a biologic discovered in laboratories. A chemical compound is a relatively smaller molecule synthesized in a laboratory, while a biologic tends to be a large molecule extracted from living bodies such as plant or animal tissues (Morrow 2004). When first discovered, this new compound or biologic is denoted as a product candidate. Medicinal or therapeutic product development is a process that starts when the product candidate is first discovered and continues until it is available to be prescribed by physicians to treat patients (Ting 2003, 2006). A compound is usually a new chemical entity synthesized by scientists from pharmaceutical companies, universities, or other research institutes. A biologic can be a protein, a part of a protein, DNA or a different form either extracted from tissues of another live body or cultured by some type of bacteria. In any case, this new compound or new biologic will have to go through the product development process before it can be used by the general patient population.

The product development process can be broadly classified into two major components: non-clinical development and clinical development. Non-clinical development includes all product testing performed outside of the human body. Non-clinical development can further be broadly divided into pharmacology, toxicology, and product formulation. In these processes, experiments are performed in laboratories or pilot plants. Observations from cells, tissues, animal bodies, or product components are collected to derive inferences for potential new products. Chemical processes are involved in formulating the new compound into drugs to be delivered into human body. In contrast, clinical development is based on experiments conducted in the human body. Clinical development can be further divided into Phases I, II, III, and IV. Clinical studies are designed to collect data from normal volunteers and subjects with the target disease in order to help understand how the human body acts on the product

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_1

candidate and, in addition, how the product candidate helps patients with the disease or leads to any potential adverse events.

A new chemical compound or a biologic can be designated as a product candidate because it demonstrates some desirable pharmacological activities in the laboratory. At the early stage of product development, the focus is mainly on evaluating cells, tissues, organs, or animal bodies. Experiments on human beings are performed after the candidate passes these early tests and looks promising. Hence non-clinical development may also be referred to as pre-clinical development since these experiments are performed before human trials.

Throughout the whole product development process, two scientific questions are constantly being addressed: (1) Does the product candidate "work"? and (2) Is it safe? Starting from the laboratory where the compound or biologic is first discovered, the candidate has to go through lots of tests to see if it demonstrates both efficacy (the candidate works) and safety. Only the candidates passing all those tests are allowed to progress to the next step of development. In the United States (US), after a drug candidate passes all the essential non-clinical tests, an Investigational New Drug (IND) document is filed with the Food and Drug Administration (FDA). After the IND is accepted, clinical trials (tests on humans) can then be performed. If this product candidate is shown to be safe and efficacious through Phases I, II, and III clinical trials, the pharmaceutical company will file a New Drug Application (NDA) to the FDA in the US. For a biologic, the application is known as Biologic License Application (BLA). If the NDA or BLA is approved, the medicinal product can then be available for general public consumption in the US. Often times the approved product is continually studied for safety and efficacy, for example, in different subpopulations or for regulatory required commitments. These post-marketing studies are generally referred as Phase IV clinical trials.

1.2 Non-clinical Development

1.2.1 Pharmacology

Pharmacology is the study of the selective biological activity of chemical substances on living matters. A substance has biological activities when, in appropriate doses, it causes a cellular response. It is selective when the response occurs in some cells but not in others. Therefore, a chemical compound or a biologic has to demonstrate these activities before it can be further developed. In the early stage of product testing, it is important to differentiate an "active" candidate from an "inactive" candidate. There are screening procedures to select these candidates. Two properties of particular interests are sensitivity and specificity. Sensitivity is the conditional probability that the screen will classify it as positive (active), given that

a compound or a biologic is truly active. Specificity is the conditional probability that the screen will call a compound or a biologic negative, (not active) given that it is truly inactive. Sensitivity and specificity involve tradeoffs; however, in the ideal case, their values are to be high and as close to one as possible.

Quantity of these pharmacological activities may be viewed as the product potency or strength. The estimation of product potency by the reactions of living organisms or their components is known as bioassay. Bioassay is defined as an experiment for estimating the potency of a drug (biologic), material, preparation, or process by means of the reaction that follows its application to living matters (Finney 1978).

1.2.2 Toxicology/Product Safety

Product safety is one of the most important concerns throughout all stages of product development. In the pre-clinical stage, product safety needs to be studied for a few different species of animals (e.g., mice, rabbits, rodents). Studies are designed to observe adverse product effects or toxic events experienced by animals treated with different doses of the product candidate. Animals are also exposed to the product candidate for various lengths of time to see whether there are adverse effects caused by cumulative dosing over time. These results are summarized and analyzed using statistical methods. When the results of animal studies indicate potentially serious adverse effects, product development is either terminated or suspended pending further investigation of the problem.

Depending on the duration of exposure to the product candidate, animal toxicity studies are classified as acute studies, subchronic studies, chronic studies, and reproductive studies (Selwyn 1988). Usually the first few studies are acute studies; that is, the animal is given one or a few doses of the product candidate. If only one dose is given to each animal during the entire study, it can also be called a single-dose study. Only those product candidates demonstrated to be safe in the single-dose studies can be progressed into multiple-dose studies. Single-dose acute studies in animals are primarily used to set the dose attempted in chronic studies. Acute studies are typically about two weeks in duration. Repeat dose studies of 30–90 days duration are called subchronic studies. Chronic studies are usually designed with more than 90 days of duration. These studies are conducted in rodents and in at least one non-rodent species. Some chronic studies may also be viewed as carcinogenicity studies since the rodent studies consider tumor incidence as an important endpoint. In addition, reproductive studies are carried out to assess the candidate's effect on fertility and conception; such studies can also be used to study product effect on the fetus and developing offspring.

1.2.3 Formulation Development

As discussed earlier, a potential new product can be either a chemical compound or a biologic. If the product candidate is a biologic, then the formulation is typically a solution which contains a high concentration of such a biologic, and the solution is injected into the subject. On the other hand, if the potential product is a chemical compound, then the formulation can be tablets, capsules, solution, patches, suspension, and other forms. There are many formulation issues that require statistical analyses. The formulation issues that stem from chemical compounds are more likely to involve widely used statistical techniques, to account for various settings in performing these experiments.

A drug is the mixture of the synthesized chemical compound (active ingredients) and other inactive ingredients designed to improve the absorption of the active ingredients. How the mixture is made depends on results of a series of experiments. Usually these experiments are performed under some physical constraints, for example, the amount of supply of raw materials, capacity of container, size and shape of the tablets. In the early stage of product development, product formulation needs to be flexible so that various dose strengths can be tested in animals and in humans. In non-clinical development stage or in an early phase of clinical trials, the product candidate is often supplied in powder form or as solutions to allow flexible dosing. By the time the product candidate progresses into late Phase I or early Phase II, then fixed dosage form such as tablets, capsules, or other formulations are more desirable.

The dose strength depends on both non-clinical evidence and clinical evidence. The product formulation group works closely with laboratory scientists, toxicologists and clinical pharmacologists to determine the possible dose strengths for each product candidate. In many cases, the originally proposed dose strengths will need to be changed depending on results obtained later from Phase II studies. These formulations are developed for clinical trial usage and are often different from the final formulation used in clinical practice. After the new product is approved for the market, final formulation should be readily available for distribution.

1.3 Pre-marketing Clinical Development

If a compound or a biologic passes the selection process from animal testing and is shown to be safe and efficacious to be tested in humans, it progresses into clinical development. In product development for human use, the major distinction between "clinical trials" and "non-clinical testing" is the experimental unit. In clinical trials, the experimental units are live human beings, whereas in "non-clinical testing" the experimental units are non-human subjects. As mentioned earlier, the results of these non-clinical studies will be used in the IND submission prior to the first

clinical trial. If there is no concern from FDA after 30 days of the IND submission, the pharmaceutical company can then start clinical testing for this drug candidate.

An IND is a document that contains all the information known about the new drug up to the time the IND is prepared. A typical IND includes the name and description of the drug (such as chemical structure and other ingredients), how the drug is processed, information about any pre-clinical experiences relating to the safety of this test drug, marketing information, and past experiences or future plans for investigating the drug both in the US. and in other countries. In addition, it also contains a description of the clinical development plan (CDP). Such a description should contain all of the informational materials to be supplied to clinical investigators, signed agreements from investigators, and the initial protocols for clinical investigation.

Clinical development is broadly divided into four phases: Phases I, II, III, and IV. Phase I trials are designed to study the short-term effects; for example, pharmacokinetics (PK, what does a human body do to the drug), pharmacodynamics (PD, what does a drug do to the human body), and the upper bound of dose range (maximally tolerated dose, MTD) for the new product. Phase II trials are designed to study the efficacy of the new product in well-defined patient populations, often with dose-response relationships as one of the major objectives. Phase III studies are usually longer-term, larger-scale studies to confirm findings established from earlier trials. These studies are also used to detect adverse effects caused by cumulative dosing. If a new product is found to be safe and efficacious from the first three phases of clinical testing, in the US, an NDA or a BLA is submitted to FDA for review. Once the new product is approved by FDA, Phase IV (post marketing) studies are planned and carried out. Many of the Phase IV study designs are dictated by the FDA to examine safety questions; some designs are used to establish new uses or indications.

1.3.1 Phase I Clinical Trials

In a Phase I PK study, the purpose is usually to understand PK properties and to estimate PK parameters (e.g., AUC, Cmax, Tmax, terms to be described in the next paragraph) of the test drug. In many cases, Phase I trials are designed to study the bioavailability of a drug or the bioequivalence among different formulations of the same drug. "Bioavailability" means "the rate and extent to which the active drug ingredient or therapeutic moiety (such as the metabolite of interest) is absorbed and becomes available at the site of drug action" (Chow and Liu 1999). Experimental units in such Phase I studies are mostly normal volunteers. Subjects recruited for these studies are generally in good health.

A bioavailability or a bioequivalence study is carried out by measuring drug concentration levels in blood or serum over time from participating subjects. These measurements are summarized into one value per subject per treatment period. These summarized data are then used for statistical analysis. Figure 1.1 presents a

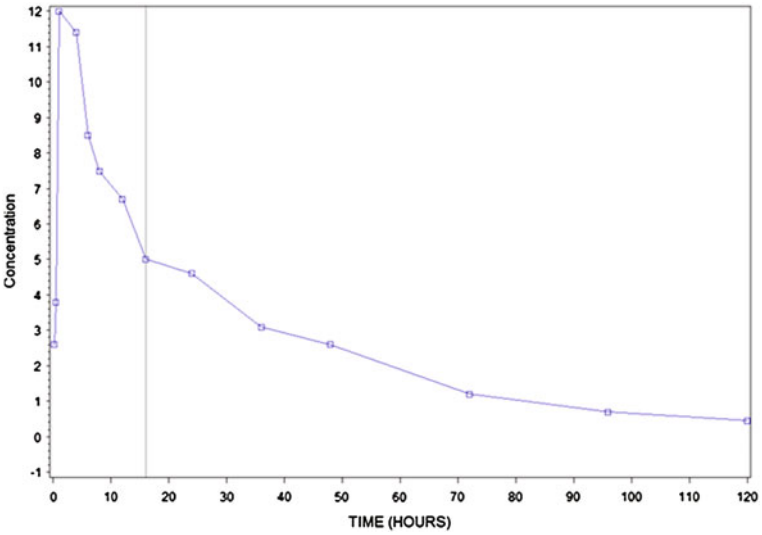

Fig. 1.1 Example of a time-concentration curve

drug concentration-time curve. Data on this curve are collected at discrete time points. Typical variables used for analysis of PK activities include area under the curve (AUC), maximum concentration (Cmax), minimum concentration (Cmin), time to maximum concentration (Tmax), and others. These variables are computed from drug concentration levels as shown in Fig. 1.1. Suppose AUC is used for analysis, then these discretely observed points are connected (for each subject under each treatment period) and the area under the curve is estimated using a trapezoidal rule. For example, AUC up to 24 h for this curve is computed by adding up the areas of the triangle between 0 and 0.25 h, the trapezoid between 0.25 and 0.5 h, and so on, and the trapezoid between 16 and 24 h. Usually the AUC and Cmax are first transformed using a natural log transformation before being included in the data analysis.

Statistical designs used in Phase I bioavailability studies are often cross-over designs where, for instance, a subject is randomized to be treated with formulation A first and then treated with formulation B after a "wash-out" period; another subject may be randomized to formulation B first and then treated with A after wash-out. In some complicated Phase I studies, two or more treatments may be designed to cross several periods for each subject. Advantages and disadvantages of cross-over designs are discussed in Chow and Liu (1999). Response variables, including AUC and Cmax, are usually analyzed using analysis of covariance (ANCOVA) models. Random effects and mixed effects linear/nonlinear models are also commonly used in the analysis for Phase I clinical studies. In certain designs, covariate terms considered in these models can be very complicated in order to assess various contributing sources of variations.

1.3.2 Phase II Clinical Trials

Phase II trials are designed to study the efficacy and safety of an experimental or investigational product. Unlike Phase I studies (except for life-threatening diseases), Phase II studies include patients with the disease that the product is developed for. Response variables considered in Phase II studies are mainly clinical efficacy variables and safety variables. For example, in a trial for the evaluation of hypertension (high blood pressure), efficacy variables are blood pressure measurements. For an anti-infective trial, response variables can be the proportion of subjects cured or time to cure for each subject. Phase II studies are mostly designed with parallel treatment groups (in contrast to cross-over treatment groups). Hence if a patient is randomized to receive dose A of the test product, then this patient is planned to be treated with dose A of this product throughout the entire study.

Phase II trials are often designed to compare one or a few doses of a test product against a placebo control. These studies are usually shorter term (e.g., several weeks) and designed with a small or moderate sample size. Patients recruited for Phase II trials are somewhat restrictive, that is, they tend to be with a certain disease severity (e.g., moderate severity), without other underlying diseases, and are not on background treatments. In practice, Phase III trials tend to relax the inclusion/exclusion criteria so that a more heterogeneous patient population can be recruited. The two major types of Phase II clinical trials are proof of concept (PoC) trials and dose-ranging trials. A typical PoC trial includes two treatment groups—placebo and a high dose of the test (experimental) product. The high dose is typically the maximal tolerated dose (MTD) or a dose that is slightly below the MTD. A dose-ranging trial usually includes multiple doses of test product and placebo. One typical dose-ranging trial uses placebo, low dose, medium dose, and high dose of the test product. Again, PoC trials and dose-ranging trials tend to be parallel designs with fixed doses. Some Phase II trials can have both the PoC part and the dose-ranging part. From a dose-ranging point of view, the objective of Phase I should be to help exploring the upper limit of dose range (MTD), and Phase II should be to help finding the lower limit of dose range. This lower limit is also known as minimum effective dose (MinED). More details of these trials will be thoroughly covered in later chapters of this book.

1.3.3 Phase III Clinical Trials

Phase III trials are longer term (can last up to a few years), larger scale (several hundreds, thousands, or even tens of thousands of patients), with less restrictive patient populations (patient characteristics are more heterogeneous), and often compared against a known active comparator (or, in some cases, placebo) for the disease to be studied. Phase III trials tend to be confirmatory trials designed to verify findings established from earlier studies.

Statistical methods used in Phase III clinical studies can be different from those used for early Phase or non-clinical studies. Statistical analyses are selected based on the distribution of the variables and the objectives of the study. Many Phase I analyses tend include descriptive, estimates and patterns observed. In Phase III, categorical data analyses are frequently used in analyzing binary data (e.g., number of subjects responded, number of subjects developed a certain condition, or number of subjects improved from "severe symptom" to "moderate symptom"). Survival analyses are commonly used in analyzing time-to-event data (e.g., time to discontinuation of the study medication, time to the first occurrence of a side effect, time to cure).

T-tests, regression analyses, analyses of variance (ANOVA), ANCOVA, and mixed effects models with repeated measures (MMRM) are useful in analyzing continuous data (e.g., blood pressure, grip strength, forced expiration volume, number of painful joints). In some cases, non-parametric analytical methods are selected because the sample size is quite limited to rely on the central limit theorem when the data do not fit any known parametric distribution well. In some other cases, the raw data are transformed (e.g., log-transformed, summarized over time) before a statistical analysis is performed in order to meet the distributional assumptions of the model. A combination of various statistical tools may sometimes be used in a product development program. Hypothesis tests are often used to compare results obtained from different treatment groups. Point estimates and confidence intervals are also frequently used to estimate subject responses to a study medication or to demonstrate equivalence between two treatment groups.

In many recent Phase III clinical development programs, enriched trial design or multi-regional clinical studies become more common. For these types of Phase III clinical trials, subgroup analyses can be useful, and these considerations will have to be incorporated at the design stage. It is important to understand that, for product approval, if a given dose that was studied in replicated Phase III studies demonstrates efficacy and safety, then that dose would be considered approvable. In order to help the regulatory agencies with a better understanding of the benefit/risk relationship of the medicinal product under study, it is helpful to incorporate the key efficacy and safety endpoints in the study. Frequently, FDA encourages sponsors to set an "End of Phase II" meeting with FDA in order to review and discuss Phase III trial designs, and possibly the entire Phase III development program.

1.3.4 Clinical Development for Products Treating Life-Threatening Diseases

In clinical development programs, concerns for products to treat life-threatening diseases can be very different from those for the development of compounds or biologics to treat other diseases/conditions. Cancer is a life-threatening disease, and

some of the concerns in developing products treating cancer are used as an example in this section. In the early stage of developing a cancer therapy, patients are recruited to trials under open label treatment with test product and some effective background cancer therapy. Traditional cancer treatments such as chemotherapy are usually cytotoxic (the therapies kills tumor cells but is also toxic to ordinary cells and tissues). Hence it is not ethical to recruit healthy normal volunteers to test the study product at Phase I. Because the experiment units are patients with cancer, it would not be ethical to use placebo control. Under these circumstances, typical Phase I cancer trials tend to be open label, dose escalation designs. These clinical studies focus on finding MTD and exploring biomarkers.

Typical oncology development programs start out with dose-escalation designs to find out MTD in Phase I. After that, the MTD is used to design Phase II trials in order to achieve proof of concept. Before entering Phase II, it is important to identify the type(s) of cancer the product will be indicated for, and patients with the targeted cancer type will be recruited to study product efficacy and safety. If the clinical or radiological response can be established from these Phase II trials, then the concept is considered proven, and the development program will continue into Phase III. Under the oncology development programs, dose-ranging studies or dose finding studies are not typically used in Phase II. The major challenge of these programs tends to be finding the right tumor type or finding the right patient population, instead of finding the right dose. Most oncology treatments are dosed at MTD for the Phase III or Phase IV trials. Therefore most of the discussions covered in this book may not necessarily be directly applicable to development programs in oncology.

Identification of biomarkers can be an important task in developing therapies for treating cancer. It is typical that cancer patients with a specific biomarker tend to respond better to a particular class of treatments. Hence certain oncology clinical development programs are planned with an enrichment strategy. An enriched patient population for a given clinical trial can be thought of as a trial recruiting patients with a specific characteristic—such as having a particular gene marker or a particular tumor type. In some cases, medicinal products for oncology are approved for the target patient population before large scale Phase III studies are completed because of unmet medical needs. When this is the case, additional clinical studies may be sponsored by National Institute of Health (NIH) or National Cancer Institute (NCI) in the US.

There are many statistical, methodological, and conceptual considerations in products developed for oncology, and these issues tend to be specific to the oncology therapeutic area (Crowley and Hoering 2012; Green and Benedetti 2012). However, Phase II clinical development programs for oncology products tend to assume that the MTD would be the target dose for clinical efficacy. Usually after the concept is proven from Phase II, such a dose would be used for Phase III development. This is different from the Phase II development programs for products treating chronic diseases. Typical product development programs for non-oncology products study a range of doses in Phase II, before engage in Phase III. Phase II

considerations in this book will include proof of concept (PoC) and dose-ranging studies. Discussions related to PoC studies could be similar between oncology and non-oncology development programs.

1.3.5 New Drug Application/Biologics License Application

When there is sufficient evidence to demonstrate that a new drug or biologics is efficacious and is safe, an NDA or BLA is compiled and submitted by the sponsor to the FDA for a decision on product approval in the US. An NDA/BLA is a huge package of documents describing all of the results obtained from both non-clinical experiments and clinical trials. A typical NDA/BLA contains sections on proposed product label, pharmacological class, foreign marketing history, chemistry, manufacturing and controls, nonclinical pharmacology and toxicology summary, human pharmacokinetics and bioavailability summary, microbiology summary, clinical data summary, results of statistical analyses, benefit/risk relationship, and others. If the sponsor intends to market the new drug in countries other than the US, then packages of documents will need to be prepared for submission to those corresponding countries, too. For example, a New Drug Submission (NDS) needs to be filed to Canadian regulatory agency and a Marketing Authorization Application (MAA) needs to be filed to the European regulatory agencies. Recently a common technical document (CTD) can now be prepared for all agencies to review. An electronic version of CTD is also known as eCTD.

An NDA/BLA package usually includes not only individual clinical study reports but also combined study results. These results may be summarized using meta-analyses or pooled data analyses on individual patient data across studies. Such analyses are performed on efficacy data to produce Summary of Clinical Efficacy (SCE, also known as Integrated Analysis of Efficacy—IAE) and on safety data to produce the Summary of Clinical Safety (SCS, also known as Integrated Analysis of Safety—IAS). These summaries are important components of an NDA/BLA. Nowadays, almost all NDA/BLA submissions are filed electronically. Electronic submissions usually include individual clinical data, programs to process these data, and software/hardware to help reviewers from FDA or other regulatory agencies in reviewing the individual data as well as the whole NDA package.

1.4 Clinical Development Plan

In the early stage of a product development program, as early as in the non-clinical stage, a clinical development plan (CDP) should be drafted. This plan should include clinical studies to be conducted in Phases I, II and III. The CDP should be guided by the draft product label. The draft label provides detailed information of how the product should be used. Hence a draft label at the early stage of product

development lays out the target profile for the product candidate. Clinical studies should be designed to obtain information that will support this given target product profile.

One of the most important aspects of labeling information is the recommended regimen for this new product. The regimen includes dosage and dose frequency. In the early stage of product development, scientists need to predict the dosage and frequency as to how the medicinal product will be labeled. Based on this prediction, the clinical development program should be designed to obtain necessary information that will support the recommended regimen. For example, if the product will be used with one fixed dose, then the CDP should propose clinical studies to help find that dose. On the other hand, if the drug will be used as titration doses, then studies need to be designed to evaluate the dose range for titration.

In addition to dose, dosing frequency deserves evaluations. Patients with chronic diseases tend to take multiple medications every day. Many patients may prefer a once-a-day (QD) drug or a twice-a-day (BID) drug. In early development of a new drug, suppose the best marketed product for the target disease is prescribed as a twice-a-day drug, and preliminary information of this test product indicates it will have to be used three or more times a day. Then the CDP needs to include studies for re-formulating the test product so that it can be used as a twice a day or even once a day drug, before it can be progressed into later phase development.

A CDP is an important document to be used during the clinical development of a new product. As a candidate progresses in the clinical development process, the CDP should be updated to reflect the most current information about the product and, depending on the most recent findings at the time, the sponsor can assess whether a new version of product label should be drafted. In case a new draft of label is needed, the development plan should be revised so that studies can be planned to support the new product label.

The overall clinical development process can be viewed in two directions as shown in Fig. 1.2. The forward direction is a general scientific process—as more data and information are accumulated, the project team knows more about the product candidate and can use this information to design subsequent studies, in order to progress the candidate along in its clinical development. Such planning, however, is based on the draft product label, which is driven by the backward direction of the process. From the draft label, there is a target product profile (TPP) for the candidate. The difference between a label claim and the TPP can be described as follows:

Claim—A statement of treatment benefit. A claim can appear in any section of a medical product's regulatory-approved labeling or in advertising and promotional labeling of prescription drugs and devices.

> *Target product (TPP)*—A clinical development program summary in the context of labeling goals where specific types of evidence (e.g., clinical trails or other source of data) are linked to the targeted labeling claims or concepts.

Depending on documented product properties found on the label, the sponsor needs to have Phase III studies to support such claims. In order to collect

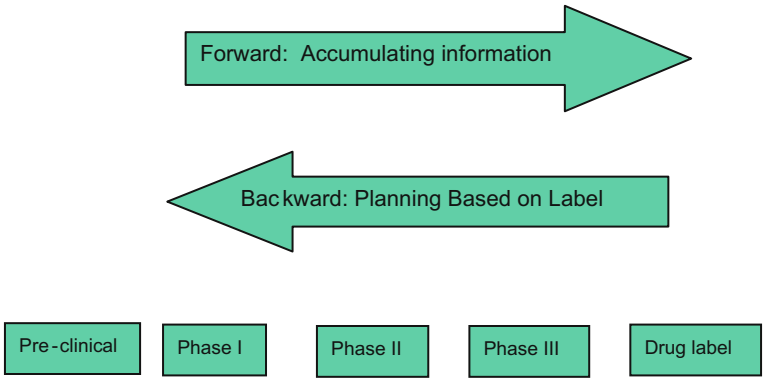

Fig. 1.2 Clinical development process

information to help design those Phase III studies, data need to be available from the corresponding Phase II or Phase I studies. Therefore the thinking process is backward by first looking at the draft label emanating from the draft target profile first and then preparing the CDP according to the draft label and supportive target profile.

1.5 Patient-Centered Outcomes

1.5.1 Clinical Outcome Assessments

A clinical outcome assessment (COA) directly or indirectly measures how patients feel or function and can be used to determine whether a treatment has demonstrated efficacy or effectiveness. COAs can also measure safety outcomes of the intervention itself or relative to another intervention. COAs measure a specific concept (i.e., what is being measured by an assessment, such as pain intensity) within a particular context of use. The context of use, which describes important boundaries or criteria under which the COA is relevant, typically includes the following elements: disease, injury, impairment or condition being treated (e.g., neuropathic pain); patient population demographics (e.g., age, disease severity, language, culture, education); and clinical treatment objectives and plan of care (e.g., reduction in pain intensity over three months with the experimental treatment administered daily) (Cappelleri and Spielberg 2015).

Of the four types of COAs (patient, clinician, non-clinician observer, and performance), the first 3 of these depend on the implementation, interpretation, and reporting from a rater on a patient's assessment—a rater being a patient, clinician, or non-clinician observer. The fourth type of a COA—performance outcomes—involves no rater judgment but is still considered a COA (and not, say, a biomarker)

because, like the other three COAs, it centers on active patient performance of a task (or activity) or on subjective assessments regarding aspects of a patient's health status. Thus, COAs involve volitionally performed tasks or subjective assessments of health status that are patient-focused or centered (Cappelleri and Spielberg 2015).

1.5.2 Patient-Reported Outcomes

Patient-reported outcome assessments have a patient as rater and rely on patient-provided responses to questions that are directly captured, without any interpretation or judgment on the part of anyone else. An example is a patient's self-report on a pain scale from 0 (no pain) to 10 (worst possible pain). Responses can be captured in several ways such as on paper, electronic forms, or by an interviewer (without interviewer's interpretation of responses).

Patient-reported outcomes are an umbrella term that includes a whole host of subjective outcomes such as pain, fatigue, depression, aspects of well-being (e.g., physical, functional, psychological), treatment satisfaction, health-related quality of life, and physical symptoms such as nausea and vomiting. While the term "health-related quality of life" (HRQoL) has been frequently used instead of PRO, HRQoL has a broad and encompassing definition that consumes a whole array of health attributes collectively, including general health, physical functioning, physical symptoms and toxicity, emotional functioning, cognitive functioning, role functioning, social well-being and functioning, sexual functioning, among others. As a more suitable term than HRQoL, PROs accommodate health questionnaires whose objective is to measure a patient's health, be it HRQoL, symptom, satisfaction with treatment, functioning, well-being—whatever the purpose—from his or her perspective rather than from a physician, caretaker, or biological measure (Cappelleri et al. 2013).

Patient-reported outcomes are often relevant in studying a variety of conditions —including pain, erectile dysfunction, fatigue, migraine, mental functioning, physical functioning, and depression—that cannot be assessed adequately without a patient's evaluation and whose key questions require patient's input on the impact of a disease or a treatment. After all, who knows better than the patient him or herself? To be useful to patients and other decision makers (e.g., clinicians, researchers, regulatory agencies, reimbursement authorities), who are stakeholders in medical care, a PRO must undergo a validation process to confirm that it is reliably and accurately measuring what it is intended to measure.

In general, the same clinical trial design principles that apply to directly assessable clinical endpoint measures (be they clinician-based or biologically-based), such as blood pressure, also apply to PROs. Although not necessarily unique to PROs, at least five characteristics tend to be associated with PROs (Fairclough 2004; Cappelleri 2013). One characteristic is that, by definition, PROs require the patient's active participation, resulting in the potential for missing data

from not only missed assessments on an entire PRO but also non-response of some items on a PRO used in a clinical study. A second characteristic is that, being subjective and not an objective endpoint like death, PROs require their measurement properties to be assessed, leading to additional steps of validation (reliability and validity) prior to their analysis on treatment effect.

A third characteristic, related to the second one of subjectivity, is that special steps and methods may be needed to enhance the interpretation of PROs. A fourth characteristic is that most PROs are multidimensional and hence produce multiple scores on various aspects of what is being measured, engendering multiple comparisons and testing of outcomes that need to be at least methodologically considered and possibly statistically addressed. The fifth characteristic is that the outcomes are generally repeated over time, calling for methods that effectively handle longitudinal data in the context of the research question. These characteristics may also pertain to other types of COA (clinician-reported outcomes, observer-reported outcomes, performance outcomes).

Identifying which components of a PRO measure (as well as other COA measures) are relevant to measuring the impact of a disease or effect of a treatment are essential to good study design and subsequent scientific scrutiny. Successful measurement of PROs begins with the development of a protocol to provide a guide for the conduct of the study. The protocol provides not only key elements of the study design but also the scientific rationale and planned analysis for the study, which are inextricably linked to study design. It is advisable to include a new or existing PRO measures (or other COA measures) as early as possible in the clinical development program.

A Phase II trial is well-suited to examine the measurement properties of a PRO measure in the targeted population. The validation process of a PRO instrument in Phase II involves whether the instrument is measuring the concept of interest (validity) and whether it is doing so dependability (reliability). Importantly, the responsiveness (within-group change) and the sensitivity (between-group change) of a PRO measure should be ascertained in Phase II before going into a confirmatory Phase III trial. Modification to and testing of an instrument should occur during the Phase II stage in order to mitigate the risk of an unacceptable instrument in Phase III, where it might be too late to make changes to the instrument. An exception occurs when the PRO instrument is already "qualified" (meaning it has been blessed a priori for use in a potential label claim by a regulatory agency) but even here it would be useful to better understand the measurement properties of the instrument in Phase II trials and Phase III trials.

For a treatment benefit to be meaningful, there should be evidence that the PRO (and COAs in general) under consideration effectively measures the particular concept (or construct or attribute) that is studied. Generally, findings measured by PROs and other COAs may be used to support claims in approved medical product labeling when the claims are derived from adequate and well-controlled investigations—Phase II trials and Phase III trials—in which PROs measure specific concepts accurately and as intended. Such PROs can be developed and assessed in accordance with regulatory guidances from the FDA (2009) and the European

Medicines Agency (EMA 2005). Even if the intent is not to seek a label claim, at least some elements from these guidance documents will be helpful for questionnaire development and assessment.

Both the EMA and the FDA have provided guidance on the qualification process for drug development tools (DDTs) (EMA 2009; FDA 2010). These guidances on DDTs can serve as a complementary adjunct to the guidance on PROs for label claims to ensure expeditious review of the new measure for use in clinical trials and potential product labeling. All too frequently, innovative and improved DDTs are delayed, which results in missed opportunities for ensuring DDTs are adequate for allowing product labeling; hopefully these new DDT processes will help with this situation.

If a DDT is qualified, analytically valid measurement on it can be relied upon to have a specific use and interpretable meaning in drug development. The objective is to develop innovative and improved DDTs that can help streamline the drug development process, improve the chances for clinical trial success, and yield more and better information about treatment or disease or both. Once a DDT is qualified for specific use, industry can use the DDT for the qualified purpose during drug development and FDA/EMA reviewers can be confident in applying the DDT for the qualified use without the need to reconfirm the DDT's utility.

Because the validation of PROs (and COAs in general) is an ongoing process, multiple protocols with each having its specific purpose may often be necessary. A protocol for a study, be it a clinical trial or a method study, should contain several essential elements: the rationale for the specific aspect of PRO being measured, explicit research objectives and endpoints, rationale for timing of assessments and non-compliance off-study rules, rationale for instrument selection, details for administration of PRO assessments to minimize bias and missing data, sample size estimation, and analytic plan (Fairclough 2004, 2010; Cappelleri et al. 2013).

A method study protocol involves by definition methodological considerations, such as which measurement properties of a PRO will be tested, and these considerations will define the design of the study. For example, if an objective is to obtain test-retest reliability data, data should be collected on at least two occasions. Contrary to a clinical trial design, which includes a pre-selected diseased population at baseline, method studies may not involve any treatment and may include a variety of subjects from healthy to severely ill for whom a PRO is designed to assess. It is advisable to have a methods study at around the same time as the Phase II program and before the Phase III program, where it may be too risky and too late to modify the instrument.

1.6 Post-marketing Clinical Development

An NDA/BLA serves as a landmark in the development of a drug or a biologic. The development process does not stop when an NDA/BLA is approved, however. Instead, objectives of the process are changed after the product is approved and is

available on the market. Studies performed after the product is approved are typically called post-marketing studies, or Phase IV studies.

One of the major objectives in post-marketing development is to establish a better safety profile for the new product. Large-scale safety surveillance studies are very common in Phase IV. Subjects/patients recruited in Phases I, II, and III are often somewhat restricted (e.g., patients would have to be within a certain age range, gender, disease severity or other restrictions). However, after the new product is approved and is available for the indicated patient population, every indicated patient with the underlying disease can be exposed to this product. Problems related to safety that have not been detected from the pre-marketing studies (Phases I, II, and III) may now be observed in this large, general patient population.

Another objective of a Phase IV trial is for the sponsor to increase the market potential for the new product by demonstrating additional benefits such as by establishing economic value and improvement in COAs such as PROs. Studies designed to achieve these objectives include humanistic (COA) studies (previously discussed in Sect. 1.5) and pharmacoeconomic studies. Such studies are often referred to as "outcomes research" studies. While these studies definitely have a place in Phase II and III, they also certainly have a placed in Phase IV.

Pharmacoeconomic studies are designed to study the direct and indirect cost or resource utilization (or both) of treating a disease (Drummond et al. 2005; Glick et al. 2005). In these studies, costs or resource utilization (e.g., hospital visits) of various FDA approved drugs are compared. Direct costs may include the price of the medication, expenses for monitoring the patient (physician's charge, costs of lab tests, and so on), costs for treating side effects caused by a treatment, hospital charges, and other items. Indirect costs may include worker loss in time or productivity and so forth. Utilities on health status obtained from Phase III trials can be used for "effectiveness" in a cost-effectiveness model, which is typically performed as part of the Phase III or Phase IV program. Cost-effectiveness analysis is used for successful market access (Muennig 2007). By showing that the new product overall costs less than another marketed product from a different company, or cost more but with a much greater benefit, the pharmaceutical or the biotech company (also known as sponsor) can increase the competitive advantage by marketing this new product.

Results obtained from "outcomes research" studies (both humanistic studies and pharmacoeconomic studies) can be used by the pharmaceutical company to promote or at least augment the new product profile. For example, if the new product is competing against another product treating the same disease or condition, the company may be able to show the new product improved the patient's quality of life beyond the improvement provided by the competing medication. The company may also be able to demonstrate that the new product brought overall savings to both the patients and the insurance carriers based on results from pharmacoeconomic studies. These studies can help expand the medical value of the new product and thereby increase the market potential for it.

Finally, another type of study frequently found during the post-marketing stage is the study designed to use the same new product for additional indications

(symptoms or diseases), expanding its use to other subpopulations of patients. A product developed for disease A may also be useful for disease B, but the pharmaceutical company may not have sufficient resources (e.g., budget, manpower) to develop the product for both indications at the same time. In this case, the sponsor may decide to develop the product for disease A to obtain approval for it to be on the market first and later develop it to treat disease B. There are also other situations that this strategy can be useful. Phase III and IV studies designed for "new indications" are quite common.

Occasionally in post-marketing studies, we may see that a product is efficacious at a lower dose than the dosage recommended in the product label. This lower dose tends to provide a better safety profile. When this is the case, product label could be changed to include the lower dose as one of the recommended doses. On the other hand, it may happen that the recommended dose may work for many patients, but the dose is not high enough for some other patients. When this is the case, a dose increase may be necessary for some patients. Based on Phase IV clinical trials, if there is a need to label a higher dose, the sponsor would seek approval from regulatory agencies to modify the product label so that a higher dose of the product can be allowed for prescription.

Patient-reported outcomes (and other COAs) have merit that go beyond satisfying regulatory requirements for a US. label claim (Doward et al. 2010). Results from Phase IV trials, as well as Phase II and III trials, can have value even when they do not appear in a label claim. Payers both in the United States and Europe, clinicians and patients themselves all have an interest in PROs that transcend a label claim for PROs and other COAs. These key stakeholders help to determine the availability, pricing, and value of medicinal products. And PROs can provide complementary and supplementary evidence distinct from other clinical outcomes in shaping the overall profile of a medical intervention. The publication of results from a well-developed PRO scale based on a well-conducted clinical study, from Phase II to Phase IV, can identify distinguishing clinical aspects, favorable and not favorable, about the treatment or disease (or both) relevant to stakeholders (Cappelleri et al. 2013).

1.7 Product Label

Throughout the entire product discovery and development process, a vast amount of data are collected and analyzed. These data were observed from molecules, cells, tissues, organs, animals, analytical laboratories, raw materials, processed batches, pilot plants, blood samples, urine samples, and human beings. All of these data are summarized, analyzed, and reported for each individual experiment or each individual clinical trial. All of the reports are then organized into integrated summary documents, and these documents then become components of an NDA or a BLA. The highest level of summary for all of these data is organized into a product label. Another way to view this is that the entire NDA or BLA could be thought of as a

pyramid where the bottom of the pyramid includes all of the raw data, the second level includes data are organized into study reports, the third level converts these reports into integrated summary documents and, eventually, the top of the pyramid contains the most condensed information, which is the product label.

A product label serves many purposes. The most important objective of a label is to communicate the product information to healthcare professionals. Physicians and other healthcare professionals need to know the efficacy and safety of a medicinal product so that a benefit/risk balance can be delivered to patients. Another important objective of a product label is to educate patients regarding how to use the product and what types of adverse events a patient may anticipate while taking this product. Hence the label serves as a communication tool to educate physicians and patients regarding the characteristics of a medicinal product.

Yet another objective of the label is to reflect the competitive advantage of the product. Usually for a given indication, there could have been several medicinal products available on the market. Hence for a new product to enter the market place, the product's sponsor needs to communicate with patients, healthcare professionals, and payers (such as insurance companies) why this new product should be used and its comparative advantage over competing products.

A label is finalized after intensive discussion between the regulatory agency and the sponsor. Note that, from a sponsor's point of view, an ideal product label would apply one single label across all over the world. In reality, however, such a situation (?) very rarely happens. This is because the fact that label negotiation is performed between the sponsor and each regulatory agency. From a regulatory point of view, the approval of any specific product means that the approved product will be marketed in the country where the agency is responsible to regulate. Since each country has its own distinctive set of circumstances—type of medical practice, products available for treating the same disease or condition, size of patient population, demographic characteristics, common co-morbidities, and so forth—most of the product labels are different at different countries based on the judgment from their regulatory bodies.

The product label must be finalized after a new product completes the pre-marketing development process but before it receives regulatory approval and is ready for indicated patient population use. Label negotiation between the sponsor and the regulatory agency takes place shortly before product approval. Nonetheless, it is advised to start planning the draft label early in the clinical development program. The draft product label evolves over time, taking several modifications and iterations, to reflect the most current understanding of the candidate under development. Eventually at the time of submission, a proposed label is included as part of the package. Regulatory agencies typically focus on the evaluation of the product itself and only after the decision is to approve the product does planning for label negotiation take priority and highest relevance.

From a sponsor's point of view, a good label should satisfy two criteria: (1) the label needs to help the product to be approvable and, then after the product is approved, (2) the label should help the clinicians to prescribe the product to help patients. Hence evidence that supports the approvability of a product can be

considered mandatory for eligible and legal market access, while evidence that supports the prescribability of a product can be considered essential for patient care. The ideal draft label should be designed with both sets of features in consideration.

Therefore, at an earlier stage of clinical development of a new product, the team needs to propose the endpoints such that a good study design will help the product to meet all regulatory requirements. Furthermore, all information obtained from clinical trials will help differentiate the product under development with its special or distinctive characteristics, especially when compared with competing medicines prescribed by the same set of physicians and other healthcare practitioners. Similarly, the dose, or range of doses, the dosing frequency and other product characteristics should all be proposed in the target product profile, which in turn will support the drug label.

A label contains clinical, non-clinical, and manufacturing information about the product. It provides necessary detailed product characteristics and describes how the product should be dosed. A complete label usually includes sections with the following titles: Description, Clinical Pharmacology, Indications and Usage, Contraindications, Warnings, Precautions, Adverse Reactions, Overdosage, Dosage and Administration, and How Supplied. The sponsor and the regulatory agency have to agree upon the content of the label. The draft label specifies the anticipated and desired characteristics of the new product. The quantitative information in a drug label is presented using descriptive and inferential statistics.

1.8 Importance of Phase II Clinical Development

Within a major sponsor (either a mid-size to large pharmaceutical company or a similar size biotech company), product candidates enter the clinical development programs on a regular basis. Many of these candidates are equipped with good potential to cure a certain disease, to largely improve a medical condition, or to slow down health deterioration. However, for a given candidate, the most crucial and difficult question is whether this particular candidate can be a successful product, or it is simply another placebo or a toxic pill. At every stage of product discovery or development, this query needs to be asked and addressed. An efficient development program is to progress a good candidate as fast as possible, so that patients will benefit from it at an earliest possible date. Meanwhile, the program should allow a bad candidate to be eliminated as early as possible so that the sponsor will not have to waste additional investments or resources on it.

Based on experiences accumulated over many decades, the best implementation of an efficient development strategy would be “design to stop.” In other words, design every study with a clear objective that if the candidate is not good enough, then stop developing it. With the concept of “design to stop” for every study design, it offers the opportunity to stop developing candidates with low likelihood of being successful at the earliest stage, saving investments and resources so that the entire development portfolio can be made more efficient. Under this strategy, a natural

question would be, if every study follows the "design to stop" principle, how about those very good candidates? In fact, if a candidate is really good, it will demonstrate its strength given a reasonable decent robust design. Hence a "design to stop" strategy tends to be one of the most efficient strategies of clinical development for new medicinal products.

One example of "design to stop" is that in most of the development programs, the first Phase II clinical trial is a PoC trial. A typical PoC trial includes two treatment groups: a test treatment group and a placebo group. The test group usually is based on the MTD of the product candidate. Such a design allows the candidate the best opportunity, if a non-decreasing dose response trend holds, to demonstrate its efficacy by using the highest allowable dose and to compare it with an agent known to be ineffective (placebo). If under this setting, the candidate still cannot show efficacy, the sponsor would be comfortable to stop developing it. This paradigm is a common example of the thinking behind "design to stop." In other words, this strategy allows an unsuccessful candidate to be eliminated at the earliest stage.

In contrast to "design to stop," another strategy would be "design to progress." If the sponsor has much confidence on the candidate being developed, and decided to design a POC study with an active control, to allow the experimental candidate to be compared with an active control intervention that is already approved for the indication being studied; the active agent could be a standard intervention and one of the market leaders. If the candidate is really strong, it could be superior to the active comparator, and the sponsor would then progress this candidate aggressively. The downside of such a design is that if the candidate is non-superior to the active control, then it is not known whether the candidate works in some way. Given this example, it helps to clarify what is meant by "design to progress," which is a contrast strategy to "design to stop." Statistically speaking, the idea of "design to stop" is implemented by a clear primary statistical hypothesis in comparing against placebo with a strict control on the level of significance (alpha). Throughout this book, the thinking process follows this "design to stop" concept.

As can be learned from the previous discussions, a first Phase II clinical trial is the earliest opportunity for a product candidate to demonstrate its efficacy in treating patients with the target disease or condition under study. The reason is that before Phase II the only data available for the product candidate are from either non-clinical studies or from healthy normal volunteers participated in Phase I trials (except for products developed to treat life-threatening diseases). Neither data could offer any opportunity to test for product efficacy in humans. Hence it becomes time critical for the sponsor to evaluate the potential of whether the candidate has a chance to be a successful medicinal product. If the likelihood of the candidate being successful is high, then the sponsor will increase the investments in developing it and in learning more features regarding this medicinal candidate. On the other hand, if the candidate is not very hopeful, then the sponsor may stop developing it or may reduce further investment on this candidate. Thus a first Phase II clinical trial tends to be a PoC study. Based on the PoC study results, a product candidate either makes it or fails to make it.

The most important deliverable of a PoC study is a clear Go/NoGo decision. As described previously, a typical PoC design is a two-treatment group parallel design. In certain situations, there could be a PoC study designed to compare the test drug against both a placebo and an active control. This is a very special case, and it will be covered in later chapters. Patients with the target disease under study are randomized into a placebo group or a test treatment group, where the dose of the test treatment tends to be the MTD of the medicinal product candidate. The null statistical hypothesis for this study is that there is no difference between the two treatment groups; the alternative hypothesis is that the test treatment is significantly superior to the placebo based on the primary measurement of the disease condition. This primary hypothesis is tested at a pre-specified Type I error rate—alpha. In this circumstance, results obtained from the PoC study is expected to deliver a clear Go/NoGo decision regarding the product candidate under development. However, in most cases, the results are not necessarily clear cut. The reasons of cases where a clear decision cannot be easily made include unclear efficacy about the primary endpoint, different responses observed from secondary endpoints, and confusing safety signals.

The two major types of Phase II clinical trials are PoC studies and dose-ranging studies. A dose-ranging design is a parallel treatment design with many doses of the test product, plus a placebo control (ICH-E4 1994). A typical dose ranging study includes a high dose, a medium dose, a low dose, and a placebo control. Patients are randomized into one of these four treatment groups, and followed to measure their responses to the assigned treatment. Data analyses are performed after the clinical data are available.

A dose-ranging study can be thought of both as a confirmatory study and as an exploratory study. It can include statistical hypotheses with multiple comparison adjustments to ensure alpha protection. It can also include modeling or estimating dose-response relationship where responses at a given dose or within a range of doses could be obtained.

One of the most important objectives of the Phase II clinical development program is to help propose particular doses or a range of doses to help with the designs in the pivotal Phase III trial. It is important to note that in product approval, for a given dose of the study product, if it is safe and efficacious at that dose, then that dose can be approvable. Hence after the product is approved and prescribed for the indicated patient population, it will be very difficult to study a dose of this product which is lower than the doses studied in Phase II or higher than a dose that is evaluated at Phase III. In other words, Phase II is the critical time of exploring a wide range of doses. If this dose range is not studied in Phase II, then the opportunity of learning dose-response relationship for that range of doses is lost. More details about these challenges will be covered in details in later chapters.

For a product candidate to enter Phase III clinical development, it is a momentous sponsor commitment to move the candidate forward. A typical Phase III program includes many larger scale, longer term clinical trials which require tens or hundreds of million dollars of investments. In the US, the FDA encourages sponsor to set an “End of Phase II” meeting with FDA. End of Phase II

and begin of Phase III is very critical from a sponsor's point of view. Because this is the last milestone point for a sponsor to stop developing a product that may not be deemed "successful" (here successful is considered as both a regulatory success and a commercial success; that is, the product can be approved and marketed so that development investments can be returned). Note that a product that is not very successful is different from a failed product. A sponsor needs to evaluate the likelihood of a candidate being successful in terms of the huge amount of investments for the entire development program. There are many examples where a candidate failed after Phase III—it did not receive any approval even when some of the Phase III trial results were beneficial and noteworthy. There are also many examples where products were successfully approved by regulatory agencies but failed to generate meaningful returns of investments. Therefore, for any given candidate at the stage between late Phase II and early Phase III, the sponsor needs to evaluate all evidence obtained from non-clinical, Phase I, and Phase II clinical studies to project how likely this candidate can be eventually successful. Based on these critical thinking and benefit/risk evaluations, a sponsor makes the very important "Go/NoGo" decision on the candidate being developed to determine whether the candidate can be progressed into Phase III.

Given the fact that Phase II is the stage where a product candidate demonstrates its activity in patients with the disease, and enable a sponsor to progress the candidate into Phase III, Phase II can be thought of as the most important phase across the entire product development process. Note that although the title of this book is about Phase II clinical development, many of the thinking process introduce in this book can be applied to Phase III, also.

1.9 Highlight of Each Chapter of This Book

Chapter 2 introduces the concept of alpha or Type 1 error, the level of significance (i.e., probability of rejecting the null hypothesis when it is true). It was the randomized controlled clinical trials that revolutionized the twentieth-century medicine. The statistical basis of approving a drug or a biologic is because of the Type I error is controlled under alpha. It helps inform the public about potential Type I error associated with approved medications. Without alpha control, it would be very difficult to demonstrate treatment efficacy delivered by a given product. In fact, the thinking of alpha control not only allows for product approval but also guides the clinical development process.

Alpha control reflects a confirmatory point of view. In other words, if the question is about decision making, such as a Go/NoGo decision, then alpha control is very important. The background on the confirmatory feature of a clinical trial relates to statistical hypothesis testing. Alpha control is also a tool used by regulatory agencies for making decisions to approve, or not to approve an application. Meanwhile, researchers typically learn from clinical trial results and the learning is generally exploratory in nature. In practice, every clinical trial has both

confirmatory and exploratory feature. Chapter 3 covers more detail about how these concepts help guide designing clinical trials.

Chapter 4 discusses considerations in designing PoC trials. Although a PoC trial looks simple with one test treatment compared against one placebo control group, there are many facets behind the design and analysis of these trials. More often than not, there are lengthy discussions after a PoC study results read out because a Go/NoGo decision is not always simple and straightforward.

If a "Go" decision is made after the PoC, the next Phase II trial would be a dose-ranging trial—to be covered in Chap. 5. A dose-ranging trial is one of the most challenging trials in clinical development programs. Most of the confusion originates from the dual features of dose-ranging studies—they are both confirmatory and exploratory. Statisticians designing these trials need to keep both features in mind. Consequences from poorly designed dose-ranging trials can be costly and time wasting.

In recent years, many sponsors attempt to combine elements of PoC and dose-ranging trials into a single study. Advantages and disadvantages of this practice are covered in Chap. 6. Many of the considerations in such a combined trial are similar to those for the individual PoC or dose-ranging trials. Nevertheless, the PoC feature of these studies can cause confusion when embedded in the combined trial. Chapter 6 introduces methods of designing these combined trials, and distinguishes the confirmatory feature based on PoC and the exploratory feature based on dose-ranging.

Inconclusiveness from a PoC study result is an important yet often poorly documented risk. Chapter 7 discusses some of the causes of this risk and how to minimize them. Usually inconclusiveness happens after the clinical data read out. However, important discussions should take place before the data are ready, after study design has been finalized. Statisticians play a key role in communicating with the team so that team members can be prepared with various scenarios from data read out. Hence a set of Go/NoGo criteria needs to be clearly documented before the blind is broken.

Chapter 8 covers data analysis of a two group PoC trial. The primary endpoint of a Phase II PoC trial is either a continuous or a discrete variable. Data analysis methods for these two types of data are introduced in this chapter.

Chapter 9 covers data analysis of dose-ranging trials when the primary endpoint is a continuous variable. The two popular types of statistical analyses for dose-ranging trials are multiple comparison procedures and modeling approaches. In the past decade, the combined method (MCP-Mod) has also received more attention. This chapter discusses situations for which analytical method is more appropriate, and presents examples to help illustrate how to use these methods. Similarly, analysis of dose-ranging trials with a binary variable as the primary endpoint is discussed in Chap. 10. As Bayesian methods have been applied broadly to various areas of statistical analyses, Chap. 11 covers the Bayesian concepts with a simple example and illustrates its application to Phase II clinical trials by continuing the dose-finding example from Chap. 9.

After Phase II is completed and that the decision is to progress the product candidate, then the next step is to design Phase III trials. Chapter 12 highlights the scope of Phase III plans in order to put Phase II trials in a larger context. It covers drug label and target product profile, non-inferiority trial designs, dose selection, drug formulations and patient populations, number of required trials for a labeling claim, number of primary efficacy endpoints, missing data issues, clinical outcome assessments, multi-regional trial issues, and the trend towards development of personalized or precision medicines.

References

Cappelleri, J. C., & Spielberg, S. P. (2015). Advances in clinical outcome assessments. *Therapeutic Innovation & Regulatory Science, 49,* 780–782.

Cappelleri, J. C., Zou, K. H., Bushmakin, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. (2013). *Patient-reported outcomes: Measurement, implementation and interpretation*. Boca Raton, Florida: Chapman & Hall/CRC Press.

Chow, S. C., & Liu, J. P. (1999). *Design and analysis of bioavailability and bioequivalence studies*. New York: Marcel Dekker, Inc.

Crowley, J., & Hoering, A. (2012). *Handbook of statistics in clinical oncology* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Doward, L.C., Gnanasakthy, A., & Baker, M.G. (2010). Patient reported outcomes: Looking beyond the claim. *Health and Quality of Life Outcomes*, *8*, 89 (Open access).

Drummon, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). New York, NY: Oxford University Press.

European Medicines Agency (EMA) & Committee for medicinal products for human use. (2005). *Reflection paper on the regulatory guidance for us of health-related quality of life (HRQOL) measures in the evaluation of medicinal products*. European Medicines Agency. www.emea.europa.eu/pdfs/human/ewp/13939104en.pdf.

European Medicines Agency (EMA) & Committee for medicinal products for human use. (2009). *Qualification of novel methodologies for drug development: Guidance to applicants*. European Medicines Agency. www.ema.europa.eu/pdfs/human/biomarkers/7289408en.pdf.

Fairclough, D. L. (2004). Patient reported outcomes as endpoints in medical research. *Statistical Methods in Medical Research, 13,* 115–138.

Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials* (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Finney, D. J. (1978). *Statistical methods in biological assay* (3rd ed.). London: Charles Griffin.

Food and Drug Administration (FDA). (2009). Guidance for industry on patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register, 74* (235), 65132–65133. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284399.htm

Food and Drug Administration (FDA). (2010). Draft guidance for industry on qualification process for drug development tools. *Federal Register, 75*(205), 65495–65496. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284399.htm

Glick, H. A., Doshi, J. A., Sonnad, S. S., & Polsky, D. (2005). *Economic evaluation in clinical trials* (2nd ed.). New York, NY: Oxford University Press.

Green, S., & Benedetti, J. (2012). *Clinical trials in oncology* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

ICH-E4. (1994). *Harmonized tripartite guideline dose-response information to support drug registration*.

Morrow, T. (2004). Defining the difference: What makes biologics unique. *Biotechnology Healthcare*, 24–29.

Muenning, P. (2007). *Cost-effectiveness analysis in health: A practical approach* (2nd ed.). San Francisco, California: Jossey-Bass.

Selwyn, M.R. (1988). Preclinical safety assessment. In K.E. Peace (Ed.), *Biopharmaceutical statistics for drug development*. New York: Marcel Dekker, Inc.

Ting, N. (2003). *Drug development, encyclopedia of biopharmaceutical statistics* (2nd ed., pp. 317–324). Marcel Dekker.

Ting, N. (2006). *Introduction and drug development process, dose finding in drug development* (pp. 1–17). New York: Springer.

Chapter 2
Concept of Alpha

2.1 Lady Tasting Tea

During a tea party in the 1920s in Cambridge, United Kingdom, the waiter presented cups, a pot of tea and milk in a tray to the party guests (Salsburg 2002). One gentleman picked up a cup, poured the tea into the cup and then added milk into the tea. At this point, a lady on the side indicated that the tea tastes better if the milk is poured in the cup first, with the tea then added to the milk. After hearing this comment, people began to discuss this suggestion. One question is whether this lady could actually tell the difference—whether tea was poured in the cup first or milk was poured in the cup first.

During the discussion, a professor suggested that he could design an experiment to check and see whether this lady could, in fact, tell the difference. This simple experiment was for the waiter to prepare 8 cups of tea, where 4 cups with tea first and 4 cups with milk first, without letting the lady know which was which. Then the lady was asked to taste each cup and to identify which cup was tea first and which was milk first. The professor's point was that by purely guessing, the lady might be able to guess 0 cups correctly, a few cups correctly, or all 8 cups correctly. The probability of such guessing could be calculated. If the probability of purely guessing was small enough, then people might be convinced that the lady could, in fact, tell the difference between tea first or milk first.

Now suppose 100 cups of tea are prepared, with 50 cups tea first and 50 cups milk first. By purely guessing, the lady may be able to guess 40 cups correctly, 45, 50, 55, or even 60 cups correctly. However, if she guessed 65 or even more cups correctly, we may be able to conclude that the lady is, indeed, able to tell the difference.

Similarly, instead of 100 cups of tea, suppose 100 patients are recruited to participate in a clinical trial, with 50 of them randomized to receive the test treatment and 50 randomized to receive the placebo treatment. Then if the sample means of these two groups are very similar, we can say that the test treatment is not

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_2

different from placebo. On the other hand, if the means of these two groups are far apart, we may be confident to conclude that the test treatment is different from placebo treatment. In a randomized, double-blind, controlled clinical trial both the subject and the investigator are blinded to treatment assignment. This type of trial is ideal to test treatment against control treatment (like placebo) in humans to see whether the test treatment is truly efficacious.

The drug regulatory agency in the United States is the Food and Drug Administration (FDA). Imagine that you are the FDA reviewer, and a drug company (a sponsor) submitted a new drug application (NDA) for you to review. How will you make the decision to approve or not to approve this application? Suppose this new drug is developed for pain reduction, and the application states that patient A took the medication and concluded that his/her pain was reduced. Patient B did the same, patient C did the same, and many more took the study drug and claimed pain reduction. Can you approve this drug for reducing pain? You cannot. All of those reports reflect the placebo effect—if a patient takes a placebo (a fake drug, a sugar pill, something we know that does not reduce pain), he or she may feel pain is reduced.

In order to avoid the approval of a placebo, the FDA requires the sponsor to demonstrate treatment effect that is statistically superior to placebo. In 1962, the Kefauver-Harris Amendment demands that an NDA would have to show that a new drug was both safe and effective for FDA to approve it. This amendment established an objective, scientific foundation for drug approval and the scientific evidence is based on statistical hypothesis testing. In order to implement this amendment, FDA had to hire statistical reviewers. As a consequence, sponsors began to hire statisticians, too. Hence the Kefauver-Harris Act is also known as "statistician employment act." This act, for demanding approved drugs to be safe and effective, in fact, has extended human life expectancy substantially. All of these scientific improvements, to some extent, originated from the Lady Tasting Tea experiment.

From this background, we see that it is nearly impossible to objectively establish the efficacy of a drug. Up to now, the best known scientific approach is to declare that the probability of this new drug being another placebo is controlled under alpha. Hence the scientific foundation for drug approval is based on this hypothesis testing frame work.

Because drug approval is based on Phase III pivotal clinical trials (ICH E-9 1998), people tend to confuse that the only confirmatory studies are Phase III trials. In fact, alpha protection based on hypothesis testing in Phase II is at least as important as in Phase III. As illustrated in later chapters of this book, one of the most critical steps in Phase II is proof of concept (PoC). If the concept is proven in Phase II, then a "Go" decision will be made for the drug candidate and it will be progressed into further development. Otherwise, if the concept is not proven, then a NoGo decision will be made and further development of this drug candidate will be stopped.

The PoC process is very similar to the FDA drug approval process—both are making a Go/NoGo decision. Although the FDA decision looks more complicated than the PoC decision, the PoC decision is, to some extent, more difficult than the

FDA approval decision. This is because that when FDA makes the decision, there are already large amount of data from non-clinical studies, manufacturing, and Phases I, II, and III clinical trials. However, for Phase II Go/NoGo decision, the results are from only one PoC clinical trial. When FDA makes a decision, the alpha protection is very critical—because a drug cannot be scientifically approved on drug efficacy but, rather, it is approved based on the Type-I error rate being protected at alpha level.

Similarly, alpha protection is also very critical in Phase II PoC studies. However, because the lack of knowledge on Phase II, sometimes PoC studies are designed with an alpha greater than 0.025 (one-sided). In the clinical development of new medicinal products, the primary interest is in demonstrating the study intervention is superior to, or non-inferior to the control agent. The focus is not to show the test treatment is inferior to control. On this basis, almost all of the efficacy clinical trials are designed with a one-sided test. The only exceptions are bioequivalence trials, or biosimilar trials, where in those cases, a two-sided statistical inference is of primary interest. However, the F-test and χ^2 test are two-sided in nature. Hence in practice, many clinical trial protocols specified that "a two-sided alpha of 0.05 is used" Under this setting, the observed *p*-value is compared with 0.05. Although the fundamental spirit of such a design is to compare half of the observed *p*-value with 0.025. In this book, in order to maintain consistency, all hypotheses tests are considered as one-sided.

For the purposes of establishing scientific evidence of a study treatment's efficacy, the null hypothesis would be that the new treatment is no different from placebo. Only after sufficient evidence exists to the contrary—that the difference of the test treatment is far from placebo—then we may conclude that the test treatment is efficacious.

For example, a treatment was studied to determine whether it lowers systolic blood pressure (SBP). A four-week clinical trial was designed to compare the SBP changes observed from the placebo group versus that from the test treatment group. Now the change in SBP from baseline (measured before a patient took the first dose of the blinded treatment) to week 4 is calculated for each patient randomized into this clinical trial. Suppose 100 patients were recruited in the trial, with 50 randomized to the test treatment and 50 to placebo. Then the group mean change in systolic blood pressure is computed from each individual patient in each treatment group.

The null hypothesis in this study is H_0: $\mu_T = \mu_P$, implying that the mean SBP change in the test treatment group is no different from the mean SBP change in the placebo group. Here μ_T denotes the mean change in SBP from the test treatment group, and μ_P denotes the mean change in SBP from the placebo group. Therefore, unless there is a sufficiently large difference in the mean changes between these two groups, we cannot conclude that the test treatment is different from placebo. In this clinical trial, where the primary endpoint is change in SBP from baseline to week 4, if the difference of subtracting the placebo mean change from the test treatment mean change is only—1.3 mm Hg, it is not likely that the test treatment is different

from placebo. Such a small treatment difference could simply be due to random chance.

However, if the mean treatment difference becomes −6 mm Hg or even −8 mm Hg, then we may consider that the test treatment could be different from placebo. When the difference is large, we may reject the null hypothesis and conclude that the alternative hypothesis can be true: the mean change in SBP from the test treatment is less than that from the placebo group. Based on the previous discussion, a one-sided alternative is considered. The statistical hypotheses can be written as follows:

$$H_0 : \mu_T = \mu_P \text{ versus } H_1 : \mu_T < \mu_P.$$

The essence of this story is that by designing and performing an experiment, we are able to convert a subjective "feeling" into an objective probability. From drinking only one cup of tea, the lady may subjectively claim whether this cup of tea tastes better or not. By repeating the tea tasting experiment for many times in a blinded fashion (i.e., without the lady knowing whether the tea or the milk was poured in the cup first), this subjective "taste" can now be converted into an objective probability statement. Similarly, after a patient is treated with a study treatment, this patient may subjectively "feel better". With a randomized, double-blind, controlled clinical trial, we are able to convert this subjective "feeling" into an objective probability statement.

The idea of modern clinical trials would not be possible without the story of lady tasting tea. In this true story, the lady was Dr. Muriel Bristol, and the professor who proposed this experiment was Sir Ronald A. Fisher. By the way, with 8 cups of tea, Dr. Bristol guessed all correctly.

2.2 Alpha Type I Error Rate

In practice, many important decisions are made without complete understanding of the entire truth. For example, a suspect may be sentenced to severe penalty while the jury does not have 100% certainty that this suspect indeed committed the crime. FDA may approve a medicinal product without 100% assurance that this product is, in fact, efficacious. Therefore, for almost every single decision based on statistical testing results, there is a risk of making an error. Table 2.1 presents the two types of errors for statistical hypothesis testing.

Table 2.1 Null hypothesis and decision

Null hypothesis	Decision	
	Accept null	Reject null
True	OK	Type I error (α)
False	Type II error (β)	OK

In the case of criminal justice, the null hypothesis is that the suspect is innocent. Unless there is sufficient evidence to demonstrate the suspect is a criminal, the null hypothesis should not be rejected. Similarly, in product evaluation, the null hypothesis is that the test product is not different from placebo. Unless there is sufficient evidence to show that the test treatment is efficacious, the null hypothesis should not be rejected. Nevertheless, if the truth is that the suspect is innocent, but the jury rejected this null hypothesis, then the jury makes a Type I error. On the other hand, if the suspect is the real criminal (null hypothesis is false), but the jury releases the suspect, then a Type II error is made. In review and evaluation of new medicinal products, if the FDA approves a product that is not efficacious, then a Type I error is made. If the FDA rejects a good product that really works, then a Type II error is made.

Alpha is the probability of making a Type I error. From an FDA point of view, alpha is the probability of approving a product when in fact this product does not work as planned (Friedman et al. 2010; Piantadosi 2005). Therefore, there should be sufficient evidence to demonstrate the product's efficacy, before it can be approved for use in the general patient population. The question becomes how to quantify "sufficient evidence." The story of lady tasting tea provided an example to quantify the level of evidence. After the randomized, double-blind, controlled clinical trials can be practically designed, conducted and reported, the evidence can be quantified by a probability (p-value) based on the null hypothesis. In other words, by controlling alpha at a small quantity considered widely acceptable, a new product can be judged as efficacious when the probability of making a Type I error is less than alpha.

The one-sided Type I error rate is commonly chosen to be 0.025, or 2.5%. Therefore, after the 1962 Kefauver-Harris Amendment, FDA approves medicinal products using this alpha level. Is alpha at 0.025 just right? Too high? Or too low? These questions are very difficult to answer.

Typically the regulatory agencies require two replicated studies in Phase III for approval, so generally the actual Type I error for any approved product is much less than 2.5% (one-sided alpha, or one in forty chance). In the case of replicate evidence with two studies, the probability of false positive reduces to one out of 1600 $[(0.025)^2 = (1/40)^2 = 1/1600]$, when these two studies are statistically independent. It has been over 50 years since the Kefauver-Harris Amendment and many important medicinal products have been approved worldwide for patients to use, the extended life expectancy and the improved quality of life brought up from these products are very real and very clear. Controlling one-sided alpha at 0.025 seems to be appropriate based on these facts. Without this concept of alpha, the medicinal product development may not be as advanced as it is today.

Although the focus of this chapter is alpha, the probability of making a Type I error, it is important to understand that Type II error is also very critical in new product development. The probability of making a Type II error is denoted as beta (β). The probability of not making a Type II error (i.e., given that the null hypothesis is false, the decision is to correctly reject the null hypothesis) is referred as the power of the test. In other words, power = $1 - \beta$. Hence when a statistician

is performing a sample size estimation using the pre-specified alpha and beta, this practice is usually referred to as "to power a study."

Both alpha and beta are important in designing a clinical trial for product development. Simply put, in terms of treatment efficacy tested by a primary efficacy endpoint, alpha is the probability that a non-efficacious treatment can be statistically tested as efficacious, and beta is the probability that an efficacious treatment can be statistically tested as non-efficacious.

2.3 Intention-to-Treat

It helps to understand how the normal (bell-shaped) curve in Fig. 2.1 is derived. Consider the example of developing a new treatment to reduce SBP. A clinical trial is designed to compare the test treatment against the placebo group with a four-week duration. One hundred patients with hypertension were randomized to test product and placebo control in a 1:1 ratio with 50 patients in each treatment group. Suppose the test product does not really work; it lowered blood pressure by only 1.3 mm Hg. After this trial is completed and 100 SBP reductions are observed, Table 2.2 captures the results of the first 8 patients: four received treatment A and four received B. Based solely on the results from these eight patients, there is a mean difference of −4 mm Hg between these two treatment groups.

Under the null hypothesis, the test treatment is not different from the placebo. Therefore, the patient response (change in SBP) should be the same (under null hypothesis) regardless of the actual treatment given. In other words, when the test treatment is not different from placebo, then whichever treatment the patient is randomized to, there should be no impact on the SBP change. Hence, under the null

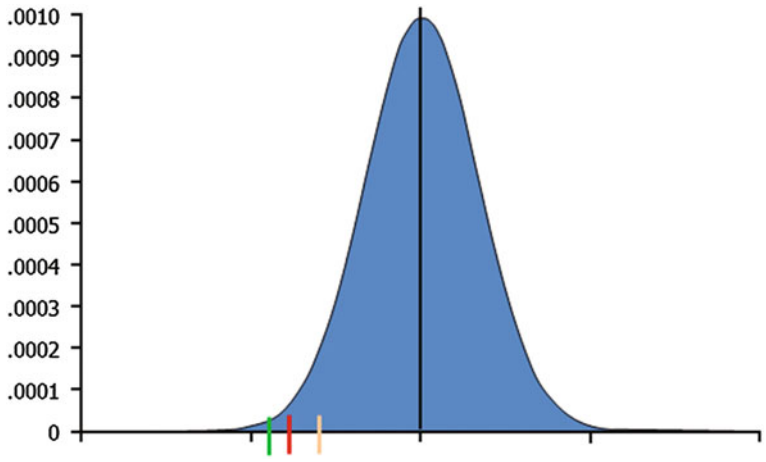

Fig. 2.1 A normal density curve with the critical value on the *left*

Table 2.2 Change in SBP from baseline to week 4 in 8 patients

	Randomization assignment	SBP reduction in clinical trial
Patient 1	A	−8
Patient 2	B	−2
Patient 3	B	−4
Patient 4	A	−7
Patient 5	A	−9
Patient 6	B	−3
Patient 7	A	−4
Patient 8	B	−3
	Mean (A-B)	−4

hypothesis, the treatment label (A or B) should not have anything to do with a patient's response, so we can permute these labels.

From Table 2.3, by permuting the labels once (under #1 of re-randomized), the mean difference between treatment A and B of the first 8 patients is 1 mm Hg. For the same reason, the labels can be permuted again—the re-randomization #2 provides a mean difference of −3.5 mm Hg. There are almost infinite number of possible permutations (under sampling with replacement) of the treatment assignments for these 100 patients, given 50 receive A and the other 50 receive B. By permuting hundreds of thousands times, the central limit theorem indicates the mean will follow a normal distribution given that the sample size is large enough (generally at least 30 subjects) and that these mean treatment differences will be expected to follow a normal distribution, like the bell-shaped curve given in Fig. 2.1, with a mean around the observed overall mean 0 based on all 100 patients together with its corresponding standard deviation.

As discussed previously, this test treatment is, in fact, not different from placebo treatment; that is, the true underlying mean treatment difference is zero (the long vertical black line in Fig. 2.1). But simply by chance, suppose that an observed

Table 2.3 Change in SBP of the 8 patients with results from 4 re-randomizations

	Randomization assignment	SBP reduction in clinical trial	Re-randomized			
			#1	#2	#3	#4
Patient 1	A	−8	B	A	A	B
Patient 2	B	−2	A	B	B	A
Patient 3	B	−4	A	B	A	B
Patient 4	A	−7	B	A	B	A
Patient 5	A	−9	A	A	B	A
Patient 6	B	−3	B	B	A	B
Patient 7	A	−4	B	B	A	B
Patient 8	B	−3	A	A	B	A
	Mean (A-B)	−4	1	−3.5	0.5	0.5

realization is −1.3 mm Hg, which falls only short of the critical point of statistical difference given by the red vertical mark (bar) in Fig. 2.1. In this case, the observed treatment difference should be the yellow vertical bar located to the right of the red vertical bar.

Nevertheless, during the conduct of this trial, if for some unknown reason, a patient randomized to treatment A was mistakenly dosed with treatment B; in this situation, the treated patients is different from the randomized patients. If the statistical analysis is performed on the treated patients, it is possible that the results become the green vertical bar to the left the red bar. Under this circumstance, the truth would be that the test treatment is not different from placebo, but the analysis based on treated patients makes it look like the test treatment is significantly different from the placebo treatment. This creates the possibility of alpha inflation. Of course, a reverse condition could also happen, which leads to a potential alpha deflation. It is critical to note that in clinical trial decision making, alpha deflation is not of a main concern. However, consequences of alpha inflation can be critical.

The concept of intention-to-treat (ITT) (Gupta 2011) is simply to follow the principle "analyzed as randomized." There are two key features to this principle: (1) all randomized patients will need to be included in the ITT analysis and (2) each patient is analyzed based on the randomized treatment, not the actual treatment taken. It is easier to understand the first principle, that all patients need to be included in the analysis. For example, if a wonder drug that cures 9999 patients out of 10,000. Unfortunately, one patient is not cured and experienced a serious adverse event (SAE). When reporting the study results, it is critical to report data from all 10,000 patients so that people reading the report get to see the whole picture. It would be biased to consider the SAE as an outlier or to delete it from the report. Therefore, every randomized patient needs to be included in the analysis and appear in the study report under the ITT principle.

The fundamental point of ITT is grounded in the permutation distribution illustrated in Fig. 2.1. Again, from the scientific-thought process, a test treatment is not considered efficacious until sufficient evidence to prove it is. Hence the null hypothesis is that the test treatment is no different from the placebo. The bell-shaped curve represents the placebo distribution. Only when the standardized observed treatment difference is very far away from the placebo mean, more than what chance will allow, then we can conclude that the test treatment is effective. In the example above, analyze as randomized would provide the true results corresponding to the given randomization.

But analysis by treatment actually received could yield a biased conclusion. Note that the green bar, in fact, reflects one of the background permutations which help construct this bell-shaped curve. Here, when the null hypothesis is in fact true, the null hypothesis would be rejected by mistake because the ITT principle was not followed. When this is the case, alpha is not protected, and the actual alpha can be artificially inflated. Of course it can be argued that if the truth is the green bar (ITT) and analyze as treated makes the yellow bar. This is the case when alpha is artificially deflated. From the statistical hypothesis testing point of view, if alpha is deflated, this is not a main concern.

2.4 Patient Analysis Sets

As discussed above, the primary analysis set for efficacy should be ITT (Gupta 2011). In practice, however, ITT may be difficult to achieve as some randomized patients may never receive study treatment may have received study treatment but have no baseline measurement, and may withdraw from the study before its completion, along with other possibilities. In application, therefore, a full analysis set (FAS) is defined for practical implementation. As indicated in ICH E-9, sponsors define FAS in the protocol and in the statistical analysis plan (SAP). Then, for statistical analysis of efficacy data, FAS is used for the primary analysis of the primary and secondary efficacy endpoints.

In reality, long before there is a clinical trial, the field of epidemiology has already been following patients over time in order to learn about the disease history and to explore the association between exposure and response (Rothman et al. 2008). One important type of epidemiological studies is the prospective cohort study. Clinical trial is simply a prospective cohort study with randomization. Based on the experiences from the past half of century, it is clear that a clinical trial is much more efficient than a (non-randomized) prospective cohort study. Hence, after randomization and blinding, patient follow up in clinical trials is the same as that performed in prospective cohort studies in epidemiology.

Nonetheless, principles used in epidemiologic cohort studies should still be followed closely in clinical trials. When reporting the results from a cohort study, it is extremely important to document all the findings truthfully—in other words, report as observed. Therefore, other than efficacy, clinical trials need to report as observed as well. The question that arises in a clinical trial is when a patient is randomized to treatment A, but actually took treatment B, how should the demographic data and the safety data be reported. From the ITT principle, it is clear that efficacy analysis uses the FAS. However, all of the patient data other than efficacy, should be reported as observed—that is, report as treated. In this case, the patient actually received treatment B, then the demographic data and the safety data of this patient should be summarized and reported with this patient in treatment B. This is because a clinical trial is still part of a prospective cohort study, and the principles of cohort studies should be followed. In protocols and SAPs for most clinical trials, a "Treated Set (TS)" is defined. ATS is very similar to an FAS. The only difference is that when a patient was randomized to treatment A, but actually took treatment B, then this patient is considered in the treatment group A in FAS but B in TS. It is important to distinguish these two concepts because analysis for efficacy is based on statistical hypothesis testing and hence needs to be alpha protected. Under this circumstance, FAS helps preserve the randomization. Patient demographic data and safety data are not alpha protected, because there is no statistical hypothesis specified for testing demographic or safety data.

Another way to look at this is that efficacy analysis is part of a designed feature of the clinical trial. On the other hand, most of the clinical trials are not designed to compare patient demography or safety. By nature of any clinical trial, product

candidate safety is an observed feature, not a designed feature. There is no statistical hypothesis behind safety reporting. Hence under the principles of cohort studies, report as observed, demographic data and safety data should be reported using the TS. This is why TS is defined, and that demographic and safety data are summarized with this patient analysis set.

Another analysis set is the per protocol set (PPS). This is a subset of FAS by excluding those patients with important protocol violations from the full analysis set. During the conduct of a clinical trial, sometimes a patient who did not meet the inclusion/exclusion criteria could be recruited to the study. In some other cases patients entered the study satisfying all of the inclusion/exclusion criteria, but they took restricted medications that may impact the efficacy response. In certain circumstances, there is a need of performing an efficacy analysis excluding these "protocol violators." When this is the case, PPS is defined at the study design stage.

When performing data analyses, statisticians have found it typical that the primary results are based on FAS, with PPS analysis performed as a sensitivity analysis or a supporting analysis. If the clinical results are robust, then the conclusion obtained from PPS should be consistent with those obtained from FAS. Yet alpha protection is based on the FAS analysis. In other words, if it is not pre-specified that the primary analysis is based on FAS, and the interpretation is based on either FAS or PPS, then alpha could be inflated. In order to reduce the likelihood that PPS results are different from FAS results, it is advised that the number of protocol violators be minimized so that PPS can be as close as possible to FAS.

It is important to note that the concept of patient analysis set is with regard to patients, but not with regard to variables observed from patients. If a patient is in FAS, but the patient's baseline measurement of the primary endpoint is missing, this patient cannot be included in the analysis of the primary endpoint (if, for instance, change from baseline is the primary endpoint). However, values corresponding to the secondary endpoint of this patient are not missing and, therefore, analysis of the secondary endpoint for this patient should be included in the FAS. Therefore, definitions of FAS, TS, and PPS for any given clinical study should be based on characteristics of individual patients, regardless of measurements from a primary or secondary efficacy variable or a safety endpoint.

2.5 Multiple Comparisons

2.5.1 Multiple Doses

For confirmatory trials, statistical procedures need to be well defined and documented before the blind is broken in order to ensure that the Type I error rate (alpha) is not arbitrarily inflated. Consider a simple example in a dose-response study with high dose and low dose of test treatment compared against placebo.

There are at least two pairwise comparisons of interest: (1) high dose vs placebo and (2) low dose versus placebo. If there is no multiple comparison adjustment and each comparison uses the entire alpha to perform the hypothesis test, then the experiment-wise Type I error rate is inflated. It is well known that in the case where there is more than one primary comparison, or more than one primary endpoint, then multiple comparison adjustment will be necessary.

Many statistical references are available in handling multiple comparison procedures (MCP) (Dmitrienko et al. 2005, 2010). In the analysis of dose-response studies, popular MCP include *Bonferroni*, *Dunnett*, *Holm*, *Hochberg*, and *gatekeeping* procedures. Some of these methods are briefly introduced here using the above simple example. Suppose a Type I error rate of alpha is pre-specified, then the experimentwise (overall study) error rate is alpha. For each comparison, there is also a comparisonwise error rate. The concept of MCP is to adjust the comparisonwise error rate so that the experimentwise error rate is protected.

The simplest adjustment is the *Bonferroni* method. Suppose there are k comparisons of interest. The Bonferroni adjustment is to divide the experimentwise error alpha by k. This reduced error rate is applied at the comparison-wise level. In the above example with two comparisons (high dose vs. placebo and low dose vs. placebo), if each pairwise comparison is performed using alpha/2, then the experiment-wise error rate is protected. Bonferroni adjustment is considered overly conservative and hence one of the less powerful methods. Nevertheless, it is easy to explain to clinical scientists, easy to implement, and can be applied without considering distributional properties of the clinical efficacy data. Hence it is preferred by several regulatory agencies.

Another simple method is to assume the k comparisons are independent, and then each pairwise comparison is tested at $1 - (1 - \text{alpha})^{(1/k)}$ level. In the above example, each pairwise comparison is performed at the $1 - (1 - \text{alpha})^{(1/2)}$ level. Although in practice, many of the pairwise comparisons are correlated, they often are positively correlated. Hence this procedure is still considered conservative. This method is somewhat more powerful than the Bonferroni adjustment and still easy to use and to interpret.

Another popular choice of MCP adjustment is the *Dunnett's* procedure (Dunnett 1955). If the data are normally distributed, then Dunnett's tests suggests to take advantage of positive correlations among these pairwise comparisons and to calculate the necessary critical points for the t-statistics. Suppose there are k pairwise comparisons with each dose against placebo, Dunnett's procedure calculates a critical value for the set of k jointly distributed t-statistics, a critical value for the $k - 1$ jointly distributed t, and so on until a set of two jointly distributed t-statistics. This MCP can be applied either in a step-down fashion or a step-up fashion.

In the case with k pairwise comparisons, the *step-down* application starts with the *largest* observed absolute t-statistic. This observed t is compared with the critical value associated with the jointly distributed t calculated based on Dunnett MCP from all k comparisons. If the observed t is less than the critical t, then stop and claim there is no dose of the study product is different from placebo; perform no further hypothesis testing. If the observed t is greater than the critical

value, then claim the particular dose associated with this largest observed absolute t is significantly different from placebo, then move on to test the next largest observed absolute t-statistic. This second largest observed t is compared with the critical value calculated from the jointly distributed t based on the set of $k - 1$ comparisons. If the observed t is less than the critical value, then claim that there is only one significant dose (obtained from the previous test) and no other dose is different from placebo. If this second largest observed t is greater than the critical value, then claim there are at least two doses being significant and move on to test the third largest observed t. Continue this sequential process until an observed t is less than the corresponding critical value, or until all pairwise comparisons are made. This is the step-down application of Dunnett MCP for when several (or even many) doses are compared against a placebo control.

As a counterpart to the step-down application, the *step-up* application of Dunnett procedure starts with the *smallest* observed absolute t. Compare this with the ordinary critical value obtained from a t-table (with one pairwise comparison). If this observed t is greater than the ordinary critical value from the t-table, then claim all k doses are significantly different from placebo and stop. Otherwise, claim the dose associated with the smallest t to be no different from placebo, and compare the second smallest observed absolute t with the Dunnett critical value calculated from the jointly distributed t based on two comparisons. If the observed t is greater than the critical value, then claim all $k - 1$ doses are significant (the dose associated with the smallest t is not significant), and stop. Otherwise, claim the first two doses are not significant and move on to test the third smallest t. Continue in this fashion until either all pairwise comparisons are performed or one of the observed t values is greater than the corresponding critical value, with the claim that dose associated with these t-statistics and all t greater than this critical t are significant.

Using the example with high dose, low dose, and placebo, the step-down application of Dunnett MCP starts with the larger observed absolute t-statistics among the two and compares it with the critical value calculated from Dunnett procedure with jointly distributed t based on all two comparisons. If the observed t is less than the critical value, claim that neither dose is different from placebo and stop. Otherwise claim the dose associated with the larger t as significant and continue to test the lower t against the critical value obtained from t-table. If the observed t is greater than the tabled t-value, then claim both doses are significant. Otherwise claim there is only one dose is significant.

On the other hand, the step-up application of Dunnett's procedure starts with the smaller observed absolute t and compares it with the critical value from the t-table. If the observed t is greater than the critical value, then claim both doses are significant and stop further comparisons. Otherwise, claim the dose associated with the smaller t is not different from placebo and continue to compare the larger observed absolute t with the critical value calculated from Dunnett's procedure with two jointly distributed t-statistics. If the observed t is less than the critical t-value, claim both doses are not different from placebo. Otherwise claim only one dose (the dose associated with the larger observed t) is significant.

Holm MCP is a step-down procedure using Bonferroni adjustment. The procedure is similar to Dunnett step-down but, instead of assuming normality and using a *t*-test, Holm's procedure uses observed *p*-values. It starts with the smallest observed *p*, and compares it with alpha/*k*. If *p* is greater than alpha/*k*, then claim none of the doses as significant, and stop. Otherwise, claim the dose with the smallest *p* is significant, and move on to test the dose associated with second smallest observed *p*-value. This *p* is compared with alpha/($k - 1$). If the second smallest *p* is greater than alpha/($k - 1$), then claim there is only one dose (the one with the smallest *p*-value) is significantly different from placebo, and none of any other doses being significant. If the second smallest *p* is less than alpha/($k - 1$), then claim two doses (both the dose with the smallest *p* and the dose with the second smallest *p*) being significant, and continue to the next smallest *p*-value. This process continues until there is a *p*-value that is not significant at that particular step or, alternatively, until all doses are compared. When compared with Dunnett's procedure, the Holm's procedure has the advantage of using *p*-values, not *t*-statistics, to make the decision, meaning that the Holm's MCP is not limited to normally distributed data. The disadvantage of the Holm's method is it being more conservative (or less powerful) than Dunnett's method, as the alpha is based on the Bonferroni adjustment.

Hochberg method is a step-up procedure. It is similar to the Holm procedure, because it uses the Bonferroni adjustment for alpha. The first step of Hochberg MCP is to test the pairwise comparison with the largest *p*-value. If it is less than alpha, then claim all *k* doses are significantly different from placebo. If not, then the next step is to compare the second largest *p*-value with alpha/2. If the observed *p* is less than alpha/2 then claim all $k - 1$ doses (all except for the dose with largest *p*) are different from placebo. Otherwise move up to check the third largest *p*-value with alpha/3. This process is continued until a *p*-value that is significant or until all doses are tested.

One popular method applied to dose-finding studies is the gate-keeping procedure. This procedure pre-specifies the order of testing and uses the entire alpha at each step. The most intuitive setting is to assume a monotonic dose-response relationship and start the first step to compare the highest dose against placebo, using the entire alpha. If the null hypothesis that "there is no difference between the highest dose of study product and placebo" is accepted, then stop and claim none of the doses is effective. Otherwise claim the highest dose is effective, and use the entire alpha to test the next highest dose against placebo. This process is repeated until either a null hypothesis is accepted, or all doses are tested. The name *gate keeping* comes from the fact that the highest dose serves as a gate; if the gate is not open, none of the following doses will be tested. The advantage of gate keeping is that this is the most powerful MCP under monotonic dose-response relationship. The disadvantage is that, at any particular dose, if the null hypothesis is accepted, then none of the lower doses would have the opportunity of being tested.

2.5.2 Multiple Endpoints

It is well known that multiplicity inflates alpha, making it more likely to falsely reject the null hypothesis of no difference. The above section discusses multiple treatment groups. Another popular multiplicity in clinical trials comes from multiple endpoints. The term endpoint could be confusing in applications. For example, in a four-week anti-hypertensive clinical trial, the "variable" of interest could be change in systolic blood pressure. The "measurements" are systolic blood pressure at each time point, and the primary endpoint is the change in SBP from baseline to week four. In scenarios like this, terms such as measurement, variable and endpoint are used loosely and interchangeably. In designing and analyzing clinical trial data, we prefer this well-defined term "endpoint."

It is common that more than one efficacy endpoint is included in a given clinical trial. When this is the case, a natural classification would be "primary endpoint" or "secondary endpoint". Under this setting, the statistical hypothesis tests applied on these endpoints form a natural hierarchy—test the primary endpoint first and then, only after the null hypothesis associated with the primary endpoint is rejected, the secondary endpoint is tested. This type of adjustment is very similar to the gate-keeping method discussed previously in dealing with multiple treatment groups. In certain other trial designs, a few "co-primary" endpoints may be introduced. The term "co-primary" implies that all endpoints specified in this group should be statistically significant in order for the study to be considered successful.

In the hierarchical specification of endpoints, typically one particular endpoint is used as the primary endpoint, and all other endpoints are considered as secondary endpoints. When more than one secondary endpoint is included in a study design, and if alpha protection is needed for secondary endpoints, it is important to pre-specify the testing order of these secondary endpoints. Again, no secondary endpoint should be tested if the null hypothesis associated with the primary endpoint was accepted. In certain clinical trials, a tertiary class of endpoints may also be pre-specified. When this is the case, these less important endpoints can be considered as "other endpoints" or "further endpoints."

Sometimes a composite endpoint is used as the primary endpoint, with components of this composite endpoint considered as secondary endpoints. One example is a time-to-event endpoint where the primary endpoint includes time to cardiovascular death, time to first stroke, and time to first severe cardiovascular event. Here the primary endpoint is time to first of any of these pre-specified events. In this example, time to each of the component event is considered a secondary endpoint.

Another example occurs with a patient-reported outcome (PRO). In the analysis of PRO measures, there is usually a set of questionnaires for each patient to answer at each visit. These questionnaires are often derived with input from patients who have the disease and the experts of the disease. There are usually recommendations as to how to score the responses to the items (questions). Usually certain scoring algorithm precede data analysis. For example, consider Western Ontario and

McMaster University (WOMAC) Osteoarthritis index (Bellamy et al. 1997), which is designed to help evaluate patients with osteoarthritis. The WOMAC total score includes a composite of 24 questions, five of these questions relate to the pain domain, two relate to the stiffness domain, and 17 relate to the function domain.

For the analysis of WOMAC data, the pain domain, stiffness domain, and function domain are analyzed separately. Typically the pain domain is considered as the primary endpoint and everything else is considered as secondary.

Suppose the objective is to show whether there is treatment difference between active treatment and placebo control in the mean change score from baseline to Week 12 for five domains on a PRO questionnaire, with the intent being a label claim. Also suppose that the researcher is considering three types of alpha or *p*-value adjustment are recommended: (1) Bonferroni method, (2) Bonferroni-Holm (Step-Down) procedure, and (3) Hochberg's (Step-Up) method.

Scenarios 1 (Bonferroni), 2 (Step-Down), and 3 (Step-Up) illustrate what may happen in the case of an instrument with five domains, the conclusion could differ among the three approaches.

Scenario 1: Illustration of Bonferroni Adjustment: Example: $K = 5$ domains with observed *p*-values of 0.20, 0.006, 0.011, 0.018, and 0.021 (i = 1, …, $K = 5$ domains).

- Bonferroni Method
  - $p(i) > \alpha/K$, then "accept" null hypothesis of no effect (α = alpha).
  - $p(i) \leq \alpha/K$, then reject null hypothesis of no effect.

Ordered *p*-values	α/K = (0.05/5)	Decision
$p(1) = 0.006$	0.01	Reject
$p(2) = 0.011$	0.01	Accept
$p(3) = 0.018$	0.01	Accept
$p(4) = 0.021$	0.01	Accept
$p(5) = 0.200$	0.01	Accept

- Conclusion: Only the domain with the lowest *p*-value (0.006) showed a treatment difference.

Scenario 2: Illustration of Step-Down Procedure: Example: $K = 5$ domains with observed *p*-values of 0.20, 0.006, 0.011, 0.018, and 0.021 (i = 1, …, $K = 5$ domains).

- Start with smallest *p*-value
- If $p(1) > \alpha/K$, then accept all null hypothesis (no treatment effect) and stop
- If $p(1) \leq \alpha/K$, then the first hypothesis [referring to $p(1)$] is rejected and then compare $p(2)$ with $\alpha/(K - 1)$
- If $p(2) > \alpha/(K - 1)$, then all remaining null hypothesis are retained

- Otherwise this second hypothesis is rejected
- Compare $p(3)$ with $\alpha/(K - 2)$ and proceed in like fashion

Ordered p-values	α/K, …, α	Decision
$p(1) = 0.006$	0.05/5 = 0.0100	Reject
$p(2) = 0.011$	0.05/4 = 0.0135	Reject
$p(3) = 0.018$	0.05/3 = 0.0167	Accept
$p(4) = 0.021$	0.05/2 = 0.0250	Accept
$p(5) = 0.200$	0.05/1 = 0.0500	Accept

- Conclusion: Only the two domains with the two lowest p-values showed a treatment difference.

Scenario 3: Illustration of Step-Up Procedure: Example: $K = 5$ domains with observed p-values of 0.20, 0.006, 0.011, 0.018, and 0.021 (i = 1, …, $K = 5$ domains).

- Step-Up: Start with largest p-value
- If $p(K) < \alpha$, then reject all null hypothesis
- If not, compare $p(K - 1)$ with $\alpha/2$
- If $p(K - 1) \leq \alpha/2$, reject all the remaining hypotheses
- Otherwise, this second null hypothesis is retained
- Compare $p(K - 2)$ with $\alpha/3$ and proceed in like fashion

Ordered p-values	α/ K,…, α	Decision
$p(1) = 0.006$	0.05/5 = 0.0100	Reject
$p(2) = 0.011$	0.05/4 = 0.0135	Reject
$p(3) = 0.018$	0.05/3 = 0.0167	Reject
$p(4) = 0.021$	0.05/2 = 0.0250	Reject
$p(5) = 0.200$	0.05/1 = 0.0500	Accept

- Conclusion: Four of the five domains showed a treatment difference.

2.5.2.1 Summary Measures or Summary Statistics

One of the methods is to use summary measures or summary statistics. For many PRO instruments, a single score can be constructed by aggregating data across different domains on the same questionnaire. Such a summary score can be used as the primary endpoint for hypothesis testing and, consequently, prevents the concern of repeated testing on multiple domains of the same instrument.

In certain situations, a combination of several steps of data transformation is necessary for analyzing multiple endpoints. One example is the O'Brien's score

(O'Brien 1984), which can be considered for all types of outcomes (be they PRO or not) with multiple endpoints embedded within them. The strategy is to combine the multiple endpoints by first ranking all the data from each variable across all patients and then summing up the ranks of all variables of interest for data analysis. Here the ranking will need to be standardized—first rank all the data from each variable and then divide the rank by the total number of subjects used in this variable. This standardization is necessary, because when some of the variables have missing data, the combination of ranks tend to assign different weights to different variables. This summed standardized rank is then used as the primary variable for data analysis. O'Brien's score is a very useful method in hypothesis testing when multiple endpoints are of equal priority. However, this method does not provide a clinically meaningful and clearly interpretable point estimate or interval estimate of the combined standardized ranks.

Summary measures can also be constructed on a particular subscale or domain of an instrument to summarize the repeated observations over time on an individual and then across individuals in the same treatment group. Examples include, for each treatment group, the average of within-subject post-treatment values, area under growth curve, and time to reach a peak or pre-specified value. The use of these summary measures begins with the construction of a summary measure for each individual, follows with the analysis of a summary measure across individuals for a within-group, and then continues with a corresponding between-group comparison. For instance, it is possible to construct summary statistics on the repeated measures within a group of individuals by taking the average rate of change over time for a treatment group and then comparing these summary statistics between groups.

A potential problem with the use of the summary score is that significant changes in some specific domains may be masked and what is really measured may become clouded or convoluted, resulting in low confidence about the effect of treatment as measured by the summary score. A major drawback of summary measures across time is that they do not fully capture the weighted and correlated nature of repeated observations on PRO measures over time.

Another way to minimize the problem of multiplicity is to restrict the number of key domains and time points, no more than a few. These key domains at specific time points should be pre-specified in the statistical analysis plan as primary endpoints for statistical inference. Other domains at other time points may be regarded as secondary endpoints. While this recommendation provides a straightforward way to handle the multiplicity issue, a major challenge is how to select the most appropriate domains and time points. One way to address this challenge is to rely on substantive knowledge, well-grounded theory, and research objectives in tandem with the nature of the disease and the intended effects of the interventions.

Often several multiple endpoints, both PRO endpoints and non-PRO endpoints, would be of clinical interest. One suitable method is to test them using a gate-keeping strategy whereby secondary endpoints are analyzed and tested inferentially in a pre-specified sequential order only after success on a primary endpoint (US Department of Health and Human Services 2009)

More generally, the key endpoints are ranked from most important to least important from the list of endpoints considered most relevant. This process can be done using a sequential method by testing additional endpoints in a defined sequence each at the usual one-sided alpha at the 0.025 level of statistical significance. The analyses cease when a failure occurs. It is important that the clinical trial protocol specify all relevant primary and secondary endpoints, and their order for inferential analysis and testing (Cappelleri et al. 2013).

2.5.3 *Other Types of Multiplicity*

Ideally, one design answers one question. When a dose-ranging clinical trial is designed, multiple comparison adjustment would be considered in order to control alpha. Sometimes an active control is used and there is a sequence of non-inferiority test, followed by a superiority test. In some cases, multiple endpoints are unavoidable, and the advice is to limit the number of endpoints. However, multiplicity may happen in many other settings. Some of the scenarios include: subgroup analyses, multiple analysis sets, time points, analytic methods, models, and interim analysis.

As mentioned earlier, it is important to pre-specify that the primary analysis will be based on the full analysis set (FAS). In a superiority trial, results obtained from analyzing the per protocol set can at most be used for the support of FAS results. In some cases, when a non-inferiority trial is designed, some may argue that PPS can be considered for the primary analysis. However, even in those cases, FAS results are still considered very important. In Phase II, most of the trials are superiority trials and hence FAS is generally used as the primary analysis set.

Longitudinal data are frequently seen in clinical trials. For example, suppose in the development of a candidate product for the treatment of hypertension, a four-week trial was designed as a Phase II study. In this clinical study, each patient is randomized to one of a few treatment groups where patients are followed on a weekly basis. At end of study, each completing patient contributes four data points (assuming no missing data), in addition to the baseline blood pressure measurements. Observations are collected at weeks 1, 2, 3, and 4. All of these patient data are included in a longitudinal data model for statistical analysis. Typical analysis would consider week 4 as the primary time point and statistical inferences are based on treatment comparisons at week 4.

After the publication of Mallinckrodt et al. (2008), mixed model with repeated measures (MMRM) became a widely accepted analytical tool for longitudinal data. Under the MMRM setting, a primary time point needs to be pre-specified, and the linear mixed effects model is then applied to the clinical data in order to obtain treatment comparisons at the primary time point. Under this setting, data observed before the primary time point are used to help with the estimation of parameters of interest. In the specification of this analytical model, it is important to focus on the primary time point. Although multiple data points are collected over time, only the primary time point will be used to formulate the statistical inference. If more than

one time point is used for treatment comparison, then alpha could potentially be inflated.

Primary analysis method needs to be pre-specified in the protocol and in the statistical analysis plan (SAP). For example, binary endpoints are frequently included in clinical trial data. Physicians are used to identify a given patient as a responder or a non-responder to the prescribed medication. Therefore it is common that in a clinical trial the primary endpoint is response to test treatment (Yes/No), and each patient is classified as one of the only two possible outcomes. For the analysis of a two-by-two contingency table (two treatment groups and two values of an outcome), there are often several statistical methods that can be used for data analysis. Widely used methods include analysis of risk difference, relative risk, odds ratio, Cochran-Mantel-Haenszel (CMH), normal approximation of binomial analysis, logistic regression, Fisher's exact test, and generalized linear models. However, for the primary analysis, only one method should be pre-specified. Other analytical methods may be used for checking the robustness of the primary results, for use in a sensitivity analysis, but only one statistical analytical method should be considered primary.

In the analysis of continuous data, typical methods include analysis of variance, regression, and analysis of covariance. For data collected over time, longitudinal analytical models could be considered. Again, only one statistical model can be pre-specified as the primary analysis model. This model needs to be clearly described so that when the statistical reviewers in regulatory agencies attempt to replicate the sponsor's analysis on the same dataset, by using the model specified in the protocol, or in the SAP, they can obtain the identical results as those presented in the clinical study report submitted by sponsors.

Model specification should clearly describe how the primary endpoint is calculated. Is it change from baseline or the post baseline outcome? Are covariates included in the model? For longitudinal data, what variance-covariance structure is assumed for the residuals over time for each subject Since only one primary model can be pre-specified, this implies that in the primary analysis of clinical trials, there will be no "model selection." This concept, again, follows the principle of "analysis respects design."

For example, a widely accepted statistical model for analysis of a continuous primary endpoint is an ANCOVA with treatment effect, stratification factor, and the continuous baseline as a covariate as follows:

$$Y_{ijk} = \mu + \tau_i + \gamma_j + \beta X_{ijk} + \varepsilon_{ijk}$$

where Y_{ijk} is the response of the kth subject from the ith treatment group with baseline stratum level j. $i = 1, \ldots, I$, $j = 1, \ldots J$, and $k = 1, \ldots, n_{ij}$, and

μ is the overall mean
τ_i is the ith treatment effect
γ_j is the jth stratum effect
β is the slope associated with the baseline measure

X_{ijk} is the baseline corresponding to the kth patient in ith treatment of kth stratum, and

ε_{ijk} is the residual of this subject

In this example, a stratification factor is included in the model as a covariate. Typically, stratification factors are accepted as potential covariates in the primary analysis model. One interesting factor could be center effect, or investigator effect. In some situations when the disease assessment is more subjective, then it is sensible to include center or investigator as a covariate in the primary model. However, if the assessment is more objective, then the team may not want to include this effect in the model because it usually takes away a number of degrees of freedom.

Much literature in biostatistics includes discussion on interim analysis, adaptive design, sample size re-estimation, and other related topics (e.g., Proschan et al. 2006). In most of these publications, the authors attempted to demonstrate that alpha is fully protected. The reason that alpha protection is a main concern is that if there is only one analysis—the final analysis—for the given clinical trial, it avoids alpha inflation. Whenever there is more than one analysis, then alpha could potentially be inflated. Hence if the protocol and the SAP pre-specified that there is no interim analysis, it will be easier to protect alpha.

Subgroup analysis is quite common in most of clinical trials. In general, statisticians are aware of the fact that these analyses are not alpha-protected. Hence, from a regulatory perspective, there should be no "statistical significance" associated with subgroup analyses. In fact, all of the alpha associated with a study is used in the primary analysis of the entire patient set. Hence, in principle, it would not be appropriate to discuss statistical significance or even provide p-values for subgroup analyses. In practice, if a p-value is reported, it should be considered merely descriptive. One impression is that the pre-specified subgroup analyses tend to be more credible than post hoc subgroup analyses. On this basis, people attempt to overly pre-specify. When this is the case, even the pre-specified subgroup analyses became lack of credibility.

Therefore, it is extremely important that methods to control all of the potential multiplicities are well defined in the SAP. The simple principle is that for a given clinical trial, there will be only one primary comparison, with only one primary endpoint, using the primary analysis set, analyzing at the primary time point, with the pre-specified analysis method, only the primary model, minimize interim analysis, and not to over interpret subgroup analysis results. This way, hopefully, alpha inflation could be reduced.

2.6 *P*-Value and Statistical Significance

Most of the popular statistical testing procedures such as z-test, t-test, F-test or chi-square test, are basically a signal-to-noise ratio. In the case of z-test or t-test, the numerator reflects the treatment difference and the denominator is the square root of

variance, adjusted by sample size. Hence the denominator represents the background noise from the data set. In a placebo controlled trial, the numerator represents the treatment difference from placebo. Therefore under the null hypothesis of no difference, the mean should be zero. If there is a strong signal (or a large numerator), then after adjustment for noise (dividing by standard error), the resulted *z*-value or *t*-value can still be very far from zero. A *p*-value can be considered as a measure of this distance—the farther away of a test statistic from the null value (zero, in this case), the smaller the *p*-value would be. Hence for a product candidate that is weak on efficacy, the signal-to-noise ratio is expected to be small and the *p*-value is expected to be large.

When the numerator degree of freedom equals one, the *F*-test is the square of the respective *t*-test, where the denominator degree of freedom of the *F*-test corresponds to the degree of freedom associated with the *t*-test. When the denominator degree of freedom of an *F*-test goes to infinity, then the *F*-test turns into a chi-square test, where the numerator degree of freedom of the *F*-test corresponds to the degree of freedom associated with the chi-square test. Both the *F*-distribution and the chi-square distribution have values that are positive. The larger these values, the smaller the corresponding *p*-values. Again, if the signal is strong after adjustment for the noise, then the corresponding *p*-value would be relatively small; if the signal is weak, then a larger *p*-value would be observed.

In the case of analysis of variance (ANOVA) or analysis of covariance (ANCOVA) for comparison of treatment effect, the *F*-test equals the mean square of treatment divided by the mean square of residual. Here the numerator is the signal, and the denominator reflects the noise (or the unexplained variability observed from the given set of data). In a clinical development program when the signal is very strong (i.e., a very good treatment which delivers strong efficacy as compared with placebo), then after adjustment for noise (by dividing the signal by the mean square of residual), the *F*-statistic can still take a large positive value and a small *p*-value could still be observed. On the other hand, if the treatment efficacy is not strong, then the corresponding numerator may not be large enough and, after dividing by the mean square of residual, the observed *p*-value could be relatively large.

Reconsider the model equation

$$Y_{ijk} = \mu + \tau_i + \gamma_j + \beta X_{ijk} + e_{ijk}$$

The analysis of variance (ANOVA) table from this model.

In this Table 2.4, SS is the sum of residuals and MS is the mean sum of residuals. Regarding the *z*-test or the *t*-test, from an extreme point of view, no matter how weak the treatment signal is, as long as the sample size is increased to a large amount, a very small *p*-value can always be observable. When this is the case, the treatment difference could be completely clinically meaningless, yet a small *p*-value is achieved. From this perspective, the focus of product development cannot be only on reducing *p*-values. It is very important to understand that a clinically meaningful treatment effect and a reasonable sample size have to be both available

Table 2.4 ANOVA

Source of variation	SS	df	MS	F
Treatment	SS_{trt}	$df_{trt} = I - 1$	$MS_{trt} = SS_{trt}/df_{trt}$	MS_{trt}/MS_R
Stratification factor	SS_{str}	$df_{str} = J - 1$	$MS_{str} = SS_{str}/df_{str}$	MS_{str}/MS_R
Baseline	SS_{bas}	$df_{bas} = 1$	$MS_{bas} = SS_{bas}/df_{bas}$	MS_{bas}/MS_R
Residual	SS_R	$d_{Rt} = df_{total} - I - J + 1$	$MS_R = SS_R/df_R$	
Total	SS_{total}	$df_{total} = \Sigma\Sigma n_{ij} - 1$		

so that a clinically meaningful effect with a statistically significant *p*-value can be delivered.

Under the ANOVA or ANCOVA setting, if the sponsor hopes to obtain a smaller *p*-value, then there are at least two ways to proceed: (1) increase the sample size (as in the situation of a *z*-test or a *t*-test), and (2) add covariates in the model which are good predictor of the outcome (dependent variable). As can be seen from the ANOVA table, the columns are source of variation, sum of squares, degrees of freedom, mean squares, and *F*; the rows are treatment, covariates, residual, and total. From the rows, every covariate explains some part of the variability (which is expressed as sum of squares). Hence as long as a covariate is not using too many degrees of freedom and explains outcome (the variation in the outcome can be explained by the covariate), then by adding covariates into the model, the mean square for residual could be reduced. Therefore, even when the treatment effect is weak, the sponsor can reduce *p*-values simply by increasing sample size or by introducing additional covariates into the ANOVA or ANCOVA model, a smaller *p*-value could be arrived. Without a strong treatment benefit, manipulation of statistical models or increasing sample sizes to reduce *p*-values are poor clinical practices that does a disservice to all stakeholders and, therefore, should be avoided as they neither help patients nor serve the sponsor's interest.

Non-statisticians in a project team tend to interpret an observed *p*-value from a two-sided test that is less than 0.05 as "statistically significant." This tendency is based on a common misunderstanding. Whether a *p*-value is statistically significant or not depends on the pre-specified alpha. In a design with two doses compared against placebo, if the pre-specified one-sided alpha is 0.025, and a Bonferroni procedure is used to adjust for multiple comparisons, then a pairwise comparison (e.g., high dose against placebo) is only statistically significant if the observed two-sided *p*-value is less than 0.0125.

In a clinical development program, it would be the statistician's responsibility to clarify this point to team members so that the term "statistical significance" is not misused or misinterpreted. In a clinical study report, there could be *p*-values supporting the primary and secondary endpoints. Sometimes *p*-values may be provided for exploratory endpoints. However, not every *p*-value is alpha-protected. In a regulatory setting for drug approval, it is very important for the statistician to clearly

articulate which *p*-values are alpha-protected (in this case, a statistical significance or non-significance can be associated with the *p*-value) and which *p*-values are not. In the case a *p*-value that is not alpha-protected, it can be considered only as descriptive or a nominal value.

2.7 Stages of a Clinical Trial

Every clinical trial can be broadly categorized into four stages—design, conduct, analysis, and report (Evans and Ting 2015). Design stage starts from the time when a clinical question is raised, or when a clinical study objective is formulated, and continues until before the first subject is screened. In most of the clinical trials, there is usually an investigator meeting, and the time of this meeting is often viewed as the end of the design stage and the starting point of the trial conduct stage. Trial conduct stage finishes at the time of clinical database lock and, after the database lock, the trial enters the analysis stage. After all analyses are ready, the medical writer works with the team to agree the key messages summarized from these analyses. Then the clinical trial enters the reporting stage.

The document to be prepared during the design stage is a clinical protocol. The protocol describes the background about the test treatment, sets the objective for the clinical trial, and covers all aspects about the study design—patient inclusion or exclusion criteria, duration of treatment, treatment groups to be included in the study, the randomization procedure, and the statistical details such as sample size calculation, definition of primary and secondary endpoints, alpha, power, and other considerations. A protocol is a comprehensive document that explicitly describe the entire clinical trial design.

The study conduct stage covers the main part of the trial. During this stage, subjects are screened and recruited to the investigators' site for assessment; clinical data are collected, reviewed, and checked. This stage can be considered as the clinical monitoring stage where most of the trial execution activities take place. Typically during this time, a SAP will be prepared and signed off. A SAP describes the statistical procedures to be applied on clinical data in performing analyses, most importantly, it describes how the alpha is protected. The SAP needs to be signed off some time before database lock.

After database lock, the clinical trial enters the analysis and the report stage. In randomized, controlled, double-blinded trials, treatment allocation is blinded to the study investigators and to the patients during the study. After last patient completes the study, and clinical data are cleaned, the database will be locked and subsequently the blind will be broken. Data analysis can only be started after blind is broken. At this time, data sets are created and statistical analyses are performed. With the analytical results being summarized into tables or figures, the statistician works with the clinician and medical writer to interpret these results. Finally, the findings from these results are written up in clinical study reports. In many studies,

these results may also be published in medical journals to deliver the messages about a product to the medical community.

Statisticians are involved in all four stages of a clinical trial, but their most important contributions to the trial team are mainly during the design stage and in the analysis/report stage. Statisticians are trained to analyze data, to interpret results, and to document statistical findings. Most of pharmaceutical statisticians are trained in design of experiments, but their graduate training in experiment design could be different from the clinical trial designs. Statistical design of experiments (DoE) tend to focus on number of factors to be controlled, interactions included in the model, and other factors. However, in clinical trial design, the most important concept is confirm and explore—confirmatory thinking is based on hypothesis testing, and exploratory practices are mostly estimations. Given this education background obtained from schools, many clinical statisticians tend to pay more attention to data analyses. However, in clinical trial design and clinical development planning, statisticians could play a much more important role. As indicated in a well-known cliché "good data analyses cannot rescue a bad design", no matter how great the statistical analysis can be performed, if the study design did not appropriately collect the necessary data for analysis, then the key question cannot be addressed.

The two major professions involve in clinical trial design are medicine and statistics. Hence it is critical for a statistician to work closely with clinicians in understanding the objective(s) of a clinical trial. The study should be designed to address the primary objective, as well as the secondary objectives. Although this collaboration is critically important, difficulties may arise in communication between these two professions in practice.

One of the key differences between these two professions originates from their educational background. Physicians are trained to care for individual patients—a physician observes and thinks about the patient coming to the clinic and tries to consider all possible angles that can best care for the patient. For a human body, many systems and organs exist, and all need to work together in order for a person to function normally. Clinicians need to understand which part of the system or organ that possibly went wrong, so that the physician can diagnose and treat the symptoms experienced by the patient. Hence medical education trained them to be thorough and complete, so that none of the details is missed in examining a patient. The focus is on the micro-level at individual patients, trying to understand what are behind the signs and symptoms observed from these individuals.

On the other hand, a statistician is trained to think about a population, and samples obtained from the given population. Given the data observed from the sample, how are inferences to be extended to the population at large. With this type of thinking, one patient is only one data point within the sample. Hence the educational background is very different between a physician and a statistician—the physician thinks about patient care and how to formulate a complete understanding of a given patient, while the statistician thinks about the population and how to make correct inferences to the population from an observed sample, which includes many patients, sometimes up to thousands or tens of thousands. In addition to educational background, when the clinician and the statistician partner on a project

team, there could be many other factors such as differences in personality, years of exposure to clinical trials, and practical experience which may affect the working relationship and mutual understanding. Therefore, clear communications between a physician and a statistician is very critical in designing clinical trials.

One major practical clinical trial design problem is that the clinical team tends to ask too many questions out of one single trial. Ideally, one design answers one question. Most of the confusion came from the lack of understanding about the priority of different questions. Often times it takes the statistician to probe these various questions and helps the team to prioritize those questions. Only after the priority becomes clear, the statistical hypothesis can then be clearly written out. If it takes a set of statistical hypotheses to address the main clinical issues, multiple comparison adjustments may need to be considered.

Analysis needs to respect design. This is an extension of the ITT principle—in other words, analyze as randomized, but also analyze as designed. No matter how well a clinical trial is designed, problems may happen during the study conduct. As indicated earlier, collecting trial data from live human beings and controlling human behavior are challenging endeavors. Hence during the study conduct many things could go wrong. Situations occurs as when patients miss medication, investigators fail to recruit, clinical data are not collected, and many other practical limitations. However, no matter what happens during the trial, by the time of performing data analysis, the statistician should respect the study design, and analyses should be consistent with the design.

2.8 Subject Selection and Choice of Alpha at Phase II

Phase III clinical trials are designed for the submission of regulatory review. These trials are also known as confirmatory trials because they are designed with an understanding of all learnings accumulated from non-clinical studies and Phases I and II clinical (human) trials. Sponsors use Phase III to confirm that all of the treatment benefits learned from previous investigations can be demonstrated to be clinically important and statistically significant. Sponsors also hope that Phase III clinical results can be generalizable so that the entire patient population with the indicated condition or disease can benefit from the product under development. Hence it is desirable that patients recruited to Phase III studies can be representative to the general patient population. This is why patient selection criteria for Phase III trials are less restrictive than in Phase II trials.

This situation in Phase III clinical trials is in contrast to the thinking of designing Phase II clinical trials. The major challenge in Phase II is proof of concept—at this stage, it is unclear whether the test product will work or not. Hence patient characteristics to be selected for trial recruitment tend to be somewhat limited and restrictive. Patients entered in Phase II studies cannot be too sick so that no matter how good the medications, these patients tangible less likely to improve. On the other hand, Phase II patients cannot be too mild so that the treatment improvement

is very difficult to detect. Thus the inclusion criteria in selecting Phase II patients have to be within a certain disease severity so that the recruited patients will provide the best opportunity to differentiate efficacy of the test product from that of the control.

Given the nature of differences between entry criteria for Phase II and Phase III clinical studies, it can be expected that the patients recruited to Phase II trials tend to be more homogeneous than patients in Phase III trials, which tend to be more heterogeneous. Accordingly, the relative efficacy from test treatment in Phase III trials tends to be diluted when compared with efficacy observed in Phase II (Chuang-Stein and Kirby 2014).

For Phase III, studies are mostly designed with a one-sided alpha of 0.025 (or equivalently, two-sided alpha of 0.05). The reason is that all regulatory agencies use this alpha level for product evaluation, and Phase III trial results are submitted to the agencies for approval decisions. However, Phase II trial results have a different audience. The objective of a Phase II study design is primarily to answer the Go/NoGo question within the sponsor. Hence the key audience of Phase II trial results is the upper management of the sponsor who has to decide whether or not to continue with the development program. The main responsibility of a trial team is to deliver the clinical data to help the decision makers within the sponsor to answer a Go/NoGo question. On this basis, the risks of making a Type I error or a Type II error are relevant and need to be taken into account by the sponsor.

Because there is a great deal of uncertainty about the treatment efficacy of the test product at early Phase II, sponsors are reluctant to make major development investments in this stage. Therefore, the sample sizes of early trials could be limited. Study sample sizes is determined by four quantities: alpha, beta, delta, and sigma. In other words, rates of Type I error (alpha), Type II error (beta), clinically meaningful treatment difference (delta), and standard deviation of patient response (sigma). In general, sigma tends to be a stable quantity. Delta is the most difficult quantity to postulate. It usually takes a lot of discussions among team members, before a reasonable delta can be proposed. Finally, rates of Type I error and Type II error can be controlled by the study team. When sample size is limited at early Phase II, and when delta cannot be increased, the only option for the team is to increase alpha or beta (or both) in order to meet the sample size limitations.

Alpha can be thought of as the probability of developing a placebo, and beta can be thought of as the probability of giving up a potentially good candidate product. At early Phase II, beta could be more important than alpha to the sponsor. Hence it is not uncommon to see early Phase II trials designed with one-sided alpha of 0.05, 0.075, or 0.1. Meanwhile, beta tends to be always no greater than 0.2. Given the variance is known, and that the treatment difference is fixed at a certain quantity, it is typical that the study team discusses how to allocate alpha or beta (or both) such that the calculated sample size is feasible.

In order to allow the best opportunity for a candidate product to demonstrate its efficacy in treating patients with the indication under study, clinical trial designs in Phase II are usually with a restricted entry criteria and often accompanied with a one-sided alpha greater than or equal to 0.025. These thinking processes allow

flexibilities in clinical trial design for early Phase II development. It can, however, also be a source of confusion for inexperienced biostatisticians. Much of this book is intended to address or clarify such points of confusion.

References

Bellamy, N., Campbell, J., Stevens, J., Pilch, L., Stewart, C., & Mahmood, Z. (1997). Validation study of a computerized version of the Western Ontario and McMaster Universities VA 3.0 osteoarthritis index. *The Journal of Rheumatology, 24,* 2413–2415.

Cappelleri, J. C., Zou, K. H., Bushmakin, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. P. (2013). *Patient-reported outcomes: Measurement, implementation and interpretation* (p. 201). Boca Raton, Florida: Chapman & Hall/CRC Press.

Chuang-Stein, C., & Kirby, S. (2014) The shrinking or disappearing observed treatment effect. *Pharmaceutical Statistics, 13*, 277–280.

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., & Offen, W. (2005). *Analysis of clinical trials using SAS: A practical guide*. Cary, NC: SAS Institute Inc.

Dmitrienko, A., Tamhane, A. C., & Bretz, F. (eds.) (2010). *Multiple testing problems in pharmaceutical statistics.* Boca Raton, FL: Chapman & Hall/CRC.

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

Evans, S., & Ting, N. (2015). *Fundamental concepts for new clinical trialist*. Boca Raton: Taylor and Francis/CRC Press.

Friedman, L. M., Furberg, C. D., & DeMets, D. L. (2010). *Fundamentals of clinical trials* (4th ed.). New York, NY: Springer.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research, 2*(3), 100–112.

International Conference on Harmonization. (ICH). (1998). E-9 statistical principles for clinical trials.

Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y., & Mancuso, J. P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal, 42,* 303–319.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics, 40,* 1079–1087.

Piantadosi, S. (2005). *Clinical trials: A methodologic perspective* (2nd ed.). Hoboken, NJ: Wiley.

Proschan, M., Lan, K. K. G., & Witttes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York: Springer.

Rothman, K. J., Lash, T. L., & Greenland, S. (2008). *Modern epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Salsburg, D. S. (2002). *Lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Henry Holt and Co.

US Department of Health and Human Services. (2009). *Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims.* http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf

Chapter 3
Confirmation and Exploration

3.1 Introduction

It should be noted that contributions of statistical science to modern medicine are not limited to statistical methods for estimating sample sizes, calculating *p*-values or confidence intervals. In addition to these methods, the indispensable and invaluable contributions are in statistical thinking and concepts and, more generally, in research methodology at are aimed at good clinical research practice. Clinical trial designs and clinical development programs are mostly based on statistical thinking and concepts. One example is the statistical hypothesis testing—the null hypothesis is that there is no difference between the test product and the placebo control.

However, unless there is strong enough evidence to demonstrate that the test product is significantly different from placebo, scientists would not reject the null hypothesis (that the test treatment is not different from the control treatment). When a decision is made that the test product is efficacious, statistical methods help quantify the Type I error, that is, the probability that the test drug is observed to be different from the control treatment when in fact it is not. This statistical thinking process establishes the foundation of designing an individual clinical trial. Furthermore, statistical thinking can also help guide the planning of an entire clinical development program composed of many clinical trials.

Another example of statistical thinking that contributes to clinical trial design is the differentiation between confirmation and exploration, the main topic of this chapter. In most Phase II and Phase III clinical trials, a study is designed with at least two purposes: decision-making and learning. Trials designed for decision-making purposes are usually called "confirmatory trials." Trials designed for learning are generally called "exploratory trials." Nevertheless, if most of these trials are used both for confirmatory purposes and for exploratory purposes, it is sensible to design studies considering both properties in a stepwise fashion.

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_3

The statistical tool to help exploration would be estimation—point estimation and interval estimation. The statistical tool to help with confirmatory is hypothesis testing.

3.2 A Motivational Example

In the Phase II clinical development of a new product, one study design may involve three treatment groups—test product, placebo, and an active control. Typically in such a study the clinical objective may be "to compare the product candidate with the active control and the placebo." Suppose the primary endpoint for this study is a continuous variable that follows a normal distribution. Then, under this objective, the project statistician is apt to propose the hypothesis

$$H_0: \mu_T = \mu_P = \mu_A, \text{ versus } H_1: \text{not all means are equal}$$

where

μ_T is the mean treatment effect of the test drug
μ_P is the mean treatment effect of the placebo and
μ_A is the mean treatment effect of the active control

Based on this hypothesis, an *F*-test may be used to compare the three treatment groups; i.e., use an *F*-test to test H_0 and, if the null hypothesis is rejected, then proceed to make the pairwise comparisons. This procedure is known as the Fisher's protected least significant difference (LSD) test.

Suppose the test product is really effective, and the pairwise comparison between the test product and the placebo is statistically significant; however, the *F*-test of the 3 groups fails to support a statistical difference. According to the Fisher's protected LSD procedure, if the overall F-test is not significant, the testing procedure stops and a pairwise test would not be performed. In this situation, a statistical procedure could delay or even stop a potentially beneficial product to reach those patients who need this medically advanced treatment. What is wrong with this proposal?

Under this circumstance, the fundamental issue is the study objective. Although the project team is interested in the comparison among all three treatment groups, there is a natural sequence of testing. The first step could be to demonstrate that the test product is superior to the placebo. This comparison will demonstrate the efficacy of the test product. Only after this step is completed, the next step examines whether the test product is non-inferior to the active control or, alternatively, superior to it. More details about this sequence of hypothesis can be found in Chap. 8.

Hence, as the study is being designed, the objectives should be explicit. The first objective is to demonstrate efficacy of the test product by comparing it with placebo, with the entire alpha applied on this comparison to protect the Type I error rate. Then the second objective is to show the non-inferiority (or superiority) of the

test product against the active control. If the sponsor is interested in making claims about the test product against the active control, then a non-inferiority margin for confidence interval estimation (for non-inferiority purpose) or a Type I error rate (for superiority) should be clearly pre-specified.

Note that if the second objective is to show non-inferiority, the Fisher's protected LSD procedure is not appropriate for this study design because part of the alternative hypothesis of interest is nested in the null hypothesis (H_0: $\mu_T = \mu_P = \mu_A$) that none of the three treatment groups are different.

A third possible pairwise comparison is between the placebo and the active control (H_0: $\mu_P = \mu_A$). Typically, this comparison is used to check for assay sensitivity. That is, if the first test fails to demonstrate the difference between the test product and placebo, then the natural question is whether the test product does not work or that the test product is in fact effective, but the study fails to indicate the treatment difference. Under this circumstance, a comparison between the active control and placebo can be used to identify where the problem is—that is, if a significant difference (between active control and placebo) is detected, then the test product is not efficacious. On the other hand, if a difference between placebo and the active control cannot be established, that implies the study may be conducted poorly, and there may still be a significant difference between test product and placebo, but this particular study fails to demonstrate that difference.

This example shows that the most important confirmatory step for this study is the pairwise comparison of test product against placebo. Only after this is confirmed, scientists can then explore the second objective. Even if both objectives are clearly pre-specified in a stepwise fashion, additional comparisons and safety profiles can still be explored from this study. The key aspect of the thinking process is that this study should be designed to confirm the primary (and possibly secondary) objective first and then move to the next step of exploration. Therefore, for each study design, it becomes a stepwise process such that seeking confirmation comes first and, only after confirmation, exploration and learning from the clinical data follow.

3.3 Clinical Development Plan (CDP)

As indicated in Chap. 1, a clinical development plan (CDP) is drafted at the very early stage, even before the product candidate enters the clinical program. In general, the early phases of clinical development (Phases I and II) are considered as exploratory development because, at this stage, it usually is unclear as to potency, dose frequency, dose regimen, and general safety profile about the product candidate. At Phase III, sponsors design and execute a large scale, longer-term trials; these are considered as confirmatory studies. Regulatory agencies review results from these Phase III studies and make decisions whether or not to approve the new product. This approach is sensible because the scientific process is to learn more first and then, based on these understandings, decisions are made. As such, a

recurring theme resonants: The process is to learn (or explore), then to confirm, and next to learn, and to confirm again, and so on.

The overall clinical development process can be viewed in two directions (Ting 2003) as shown in Fig. 1.2. One is the forward scientific process—as more data and information accumulate, more is known about the product candidate in order to design later phase studies to help progress the candidate in drug development. On the other hand, the planning is based on the draft product label. Depending on the product properties to be demonstrated on the label, the thinking process is formulated with the goal in mind by looking at the target profile first and then crafting the CDP according to the draft label motivated by the target profile. Although the product development process is a first learning (exploratory) and then confirming process, study design for each individual trial should be thought of in the reverse way. A clinical trial should be designed with the confirmatory objective in mind first and, after the study can be assured to achieve that goal, additional properties of the study product can be explored.

According to the "design to stop" strategy, the most efficient development strategy is to allow each individual clinical trial before Phase III to have the opportunity to stop the product candidate for further development. As indicated earlier, although the focus of this book is about Phase II, many of the principles covered in this book can be applicable to Phase III, also From this point of view, every such individual clinical trial can in principle be considered as a Go/NoGo trial. Before engagement in Phase III, the most important question of every clinical trial should be a Go/NoGo decision. For many (if not most) clinical development project teams, members tend to not focus on this question because, buoyed from optimism after completing a particular trial, they may be more inclined to move to the natural next step of designing a new study and moving forward. This scenario is the typical case because an implied "Go" decision has been made based on favorable (or, at least, not unfavorable) study results observed from this given trial. For those candidates that are promising or somewhat promising, this could be a somewhat easy decision when the study results are available. If there is no critical issue from these clinical data, the team tends to make a Go decision.

However, the difficulty comes from the design stage which occurs before any team member sees the data. The key question is, how to design a clinical trial so that by the time when blind is broken and data are available, an easy decision can be made? A good design tends to address the Go/NoGo question directly. Thus, from the view point of planning for the entire development program, it would be more efficient to build in a Go/NoGo criteria in order to correspond to every study planned in the program through Phase II. In other words, a critical evaluation of the Go/NoGo criteria that should be built in for every Phase II clinical trial before a product candidate enters Phase III.

3.4 Clinical Study Design and Sample Size Calculations

For every clinical trial, before the study design, there is always a clinical question of interest. Based on this clinical question, sample sizes are calculated. The 4 quantities necessary for sample size calculations are alpha, beta, delta and sigma (α, β, δ, and σ; see Table 3.1).

Sample size calculations are important because this number reflects the assumptions a project team is willing to make. These assumptions include selection of primary endpoint, tolerability of errors (Type I and Type II errors), anticipation of the clinical difference to be observed, and the budget to execute the study. Each study design is a compromise of these assumptions. Selection of the primary endpoint is based on the understanding of the disease to be studied, regulatory guidelines, mechanism of action of the experimental intervention, study duration, and how difficult it is to reach clinically meaningful efficacy.

Based on all of these considerations, a sample size is calculated with the expectation that the clinical data will help make a Go/NoGo decision on the product candidate. Hence this part is the confirmatory part of the study design. After the sample size is determined, the clinical study is designed to collect additional efficacy and safety data for scientists to learn more about the study product. While patient-reported outcomes (PROs) have been primary endpoints in several therapeutic areas such as sexual dysfunction, pain and rheumatology, PROs have been also secondary endpoints in many other circumstances (e.g., studies on oncology, smoking cessation, cardiovascular, and diabetes) ti help assess health status such as physical and mental functioning and well-being.

In many Phase II clinical trials, the secondary endpoints such as PROs are not alpha-protected. For example, in many oncology trials, PROs are secondary endpoints while the progression-free survival is the primary endpoint. Hence these PRO endpoints are not typically considered in the sample size calculation. However, in Phase III, some of these endpoints may be considered to be included in the product label. In this case, sample size for the given Phase III trial should be sufficiently large so that the secondary endpoints (including PROs if they are secondary outcomes) can be confirmed to support the label statements. Additional endpoints and other design features for this study could be for exploratory purposes.

Table 3.1 Factors considered for sample size calculations

Symbol	Meaning	Typical values	Implications
α	Type I error rate	One-sided at 0.025	Develop a placebo
β	Type II error rate	0.1 or 0.2	Give up a good drug
δ	Clinically meaningful difference	Obtained from regulatory guidance, literature, etc.	Achievable? Regulatory acceptable?
σ	Standard deviation	Obtained from literature or pilot studies	Mis-specification of σ could affect sample size calculation

Here the two-step approach of designing a clinical study is to first use the primary study objective to determine the sample size, which serves as the confirmatory step; the second step is to include the necessary adjunct or exploratory pieces into the study design.

3.5 Statistical Analysis Plan (SAP)

For most Phase II/III clinical trials, a SAP exists for each study. In contrast to the CDP, which guides the entire development program, a SAP provides guidance of how to analyze data collected from a single clinical trial. The SAP is drafted around the time a study protocol is being finalized. It needs to be completed and signed off before the blind is broken. This document serves as a contract between the statistician and the project or the trial team. It can also be submitted to regulatory agencies as a supporting document illustrating that all statistical analyses are pre-specified and prospectively defined, not based on post hoc data dredging.

The SAP describes the statistical analysis methods in more details than in the protocol. Usually topics covered in a SAP include definitions of the primary endpoint, the primary comparison, the primary analytical population, the primary analysis method, the primary time point, and all secondary analyses. The SAP might also specify the safety analysis and other relevant analyses [e.g., pharmacokinetics (PK) analysis, pharmacogenomics analysis, patient-reported outcomes, additional doses, and others].

With these detailed specifications, the statistician collaborates with the project (or the trial) team to agree on what exactly can be expected when the study results are ready. The team may also request for specific analyses to address particular clinical issues. Thus a SAP serves as a contract with the team. The main purpose of a SAP is to ensure that the Type I error is not inflated. All this effort is made for a confirmatory purpose—when the study results demonstrate drug efficacy, the team is certain that by progressing this candidate product, the probability of developing a placebo is controlled to be less than the stated level of significance (alpha). After the primary analysis is performed, and the candidate product is shown to be efficacious, the team then explores other characteristics of this product candidate (secondary efficacy analyses, product safety, and other issues of interest).

As covered in Chap. 2, for regulatory purposes, multiplicities could potentially inflate alpha—multiple comparisons, multiple endpoints, multiple analysis sets, and other multiplicities—should be well-controlled so that the pre-specified experiment-wise Type I error rate (alpha) can be fully protected. Of course a simple design with only one primary comparison, one primary endpoint, one primary analysis set, one primary time point, one primary analysis method with one primary model without any interim analysis would clearly help with alpha protection. Hence in a proof of concept study where the key question is a Go/NoGo decision, all of these simplicities help with alpha protection, with the intention that the decision will not be too difficult after study results are ready. As also mentioned

earlier, the most efficient clinical development strategy is to make every pre-Phase III clinical study a Go/NoGo study. From this point of view, therefore, a very simple design that answers one single key question could be one of the most efficient strategies in clinical development of new medicinal products.

However, when multiplicities exist, statistical adjustment for handling multiplicity should be included in both the protocol and the SAP. It should be noted that *p*-values used to test all of the alpha-protected statistical hypotheses can be interpreted; that is, a "statistical significance" or "statistically non-significant finding" can be associated with those *p*-value(s). A preferred way to move forward would be that, other than those hypotheses where alpha is protected, no additional *p*-value will be provided in a study report, assuming that all those statistical summaries are for exploratory purposes only. The differentiation of confirmation and exploration in this case would be that all alpha-protected statistical hypotheses tests serve confirmatory purposes; all of the additional results are considered exploratory from a regulatory perspective. Exploratory analyses include all those efficacy analyses that are not alpha-protected—background and demographic data, safety summaries, outcome research clinical outcomes (including PROs, if applicable, and resource utilization metrics), and others. Again, the study design thinking is to confirm first and then to explore.

3.6 Application Example—Another Three Group Phase III Design

In the Phase III clinical development of valdecoxib, one of the dental pain studies includes 3 treatment groups: valdecoxib compared with placebo and rofecoxib (Fricke et al. 2002). In this study, valdecoxib is the test drug and rofexocib is the active control. The primary purpose of this study is to demonstrate that valdecoxib is superior to rofecoxib. However, in order to obtain drug approval, it is necessary to also include a placebo group. As a result, a study becomes a three-group Phase III study for dental pain.

Dental pain is a well-established therapeutic indication. Assume that the primary endpoint, clinically meaningful difference, and variability of clinical measures are all established from previous studies. Hence from a product approval perspective for a dental pain study to demonstrate superiority to placebo, there is generally no need to have more than 50 patients per treatment group.

For this Phase III dental pain study, the primary endpoint is the proportion of patients experiencing treatment failure, defined as those patients who took rescue medication within a given period of time. Since the primary objective is to demonstrate superiority of the test drug over the active comparator, this study should be powered based on this primary comparison. Because the efficacy margin for superiority to an active control is less than the margin to the placebo, the sample size calculation indicated 80 patients per treatment group in order to meet the

pre-specified power. However, the secondary objective of this study requires the test drug to be superior to placebo. Hence a placebo group with 40 patients would be sufficient. Based on these considerations, this dental pain study is designed with an unbalanced randomization of 2:2:1 ratio of the three treatment groups—80 patients for test drug and for the active comparator and 40 patients for the placebo group.

In this example, the most important confirmatory step is to demonstrate the superiority of valdecoxib to the active control (rofexocib). The next confirmatory step is to show that valdecoxib is superior to placebo. After these two confirmatory steps, investigators are interested in exploring other characteristics of valdecoxib. The key to this imbalanced design is to recognize that the first confirmatory step requires a larger sample size than the second step and, to that end, a 2:2:1 randomization ratio introduces additional efficiency to this Phase III study design.

3.7 Application Example—Dose Selection

In a typical dose-ranging clinical trial, a placebo and a few active test doses are studied. One frequently used dose-ranging study is designed with four treatment groups: placebo, low dose, medium dose, and high dose of the test drug. Clinical data obtained from these studies can be viewed from two different angles. First, the statistician can perform pairwise comparisons to see if any of the doses is superior to placebo. Second, the statistician can consider parametric model that includes all doses to describe a dose-response relationship. When analyzing data from dose-ranging studies, statisticians consider two families of common statistical methods: multiple comparison procedures (Tamhane et al. 1996) and modeling approach (Ruberg 1995).

As discussed in Chap. 2, multiple comparison procedure (MCP) is usually performed with pairwise comparisons of each test dose against placebo. When pairwise comparisons are made, the level of significance (alpha, the probability of making a Type I error) could be inflated and multiple comparison adjustment need to be made. For example, in the case of a four treatment group dose-ranging design, the three pairwise comparisons are high dose versus placebo, medium dose versus placebo, and low dose versus placebo. The Bonferroni adjustment divides alpha by 3, and each pairwise comparison is tested at the alpha/3 level. Other MCP adjustments include Dunnett's method, Holm's procedure, closed testing, and others. One of the frequently used procedures in drug development is the gate keeping adjustment (Dmitrienko et al. 2006).

As indicated in Chap. 2, an example of the gate-keeping procedure in a four treatment group dose-response study can be described as follows. The first step is to test the high dose against placebo at alpha. If the null hypothesis is accepted, then stop and claim that none of the test dose is effective. If a p-value less than alpha is observed, then claim that the high dose is significantly different from placebo and test the medium dose versus placebo at alpha. If a p-value is greater than or equal to

alpha, then accept null hypothesis and stop; claim that only the high dose is effective and that none of the lower doses are effective. If the *p*-value from the second step is less than alpha, then claim that both the high dose and the medium dose are effective and move to test low dose versus placebo at level alpha.

In general, MCP is very useful in protecting alpha so that it is not arbitrarily inflated. Controlling alpha is extremely important in Phase II and Phase III studies where regulatory agency will need to make a decision about the study drug. The other family of methods in analyzing dose-ranging clinical data is based on the dose-response modeling approach. A simple example is to assume that the three test doses and placebo forms a linear dose-response relationship. Based on such a model, this four-treatment group study can be analyzed using a linear contrast test, or a simple linear regression, using dose as the regressor. Of course, other models such as Emax, logistic, and many others can also be considered for data analysis. Results obtained from the analysis of a linear dose-response model will help investigators to determine whether or not there is a linear dose-response relationship. Also the coefficient estimated from this model can be used to study various properties of the dose-response relationship. The advantages of using a dose-response modeling approach include at least the following five characteristics: (1) helps to identify a target dose of interest [e.g., minimum effective dose, ED_{90} (a dose with 90% of maximum efficacy)]; (2) provides confidence intervals on selected doses; (3) avoids multiple testing concerns; (4) helps better understanding of dose-response relationship; and (5) guides planning of future studies. With identification of a target dose of interest, doses do not necessarily have to be the doses being studied (a dose-response model helps interpolate responses from observed doses to estimate the responses of doses that were not selected and observed).

Pinheiro et al. (2003) propose a method to combine the MCP and the modeling approaches, denoted as the MCP-Mod method, for analyzing dose-ranging studies. The basic idea is to consider a set of candidate models at the design stage and, then for each model, construct an optimal contrast corresponding to each model. If there is a single candidate model with only a given set of parameters, then the data analysis will be performed with just one contrast. Hence no multiple comparison adjustment is necessary.

Once a dose-response model is determined, this model can be used to estimate target doses of interest and construct confidence intervals on efficacy responses of these doses. Information obtained from this model can help to design other Phase II studies and even Phase III studies. The MCP-Mod is an integrated approach to explore dose-response relationships, while confirming that the test product is efficacious based on hypothesis testing. Therefore, in the desing of a dose-response study using MCP-Mod approach, the first step is to confirm there is a dose response that fits at least one of the candidate models, and this can be accomplished by the "MCP" step. Then, after the confirmation of dose response, investigators continue to explore using dose response modeling, the "Mod" step. This approach can be considered as an application of the paradigm on confirm–then-explore.

Table 3.2 Pairwise comparisons to placebo

Parameter	Estimate	Standard error	*t*-value	One-sided *p*-value	Marginal 90% Conf. Int.
$\mu_1 - \mu_0$	0.6038	0.2253	2.68	0.0044	(0.2296, 0.9780)
$\mu_{0.6} - \mu_0$	0.5895	0.2253	2.62	0.0052	(0.2153, 0.9638)
$\mu_{0.2} - \mu_0$	0.4654	0.2253	2.07	0.0201	(0.0912, 0.8396)
$\mu_{0.6} - \mu_0$	0.1118	0.2253	0.50	0.3103	(−0.2624, 0.4861)

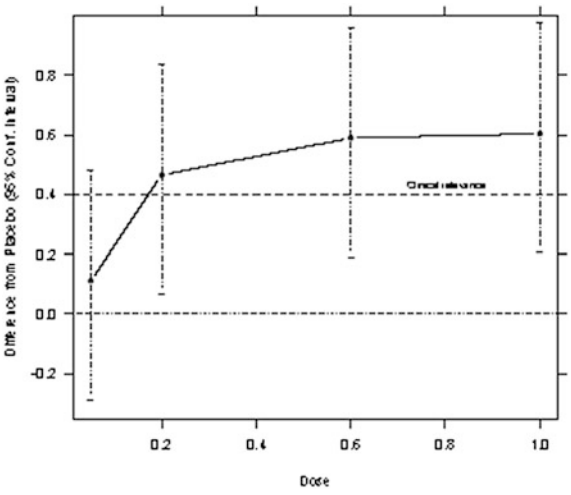

Fig. 3.1 Estimated treatment differences with respect to placebo and associated marginal 90% confidence intervals for each dose in the Phase II example

One example can be found from Westfall and Krishen (2001). Table 3.2 displays the pairwise comparisons of treatment to placebo using a 5% one-sided alpha. These data are also plotted in Fig. 3.1. Pinheiro et al. (2003) applied the Emax, linear in log dose, linear, exponential and quadratic (umbrella shape) models to re-analyze data in this example, and demonstrate how the MCP-Mod method can be used to estimate the minimum effective dose.

3.8 Proof of Concept and Dose Ranging

In the development of a new medicinal product for non-life threatening diseases, a proof of concept (PoC) study is designed at early Phase II to evaluate whether the test treatment could show product efficacy based on the non-clinical and early clinical findings. Typical Phase I programs for these drug candidates recruit healthy normal volunteers to help establish the pharmacokinetics/ pharmacodynamics (PK/PD) profile. But drug efficacy can only be studied in patients with the target disease. For these products, therefore, no efficacy data is available at the beginning

of Phase II. In many situations, the drug candidate may show strong efficacy from animal models and non-clinical experiments but may fail to provide efficacy to human subjects. Hence a PoC study is very critical in product development to help sponsors make a decision as to whether or not to follow-up developing this candidate.

A typical PoC study is designed with two treatment groups—a high dose of the study product and a placebo group. The high dose is usually the maximal tolerated dose (MTD) or a dose that is a bit lower than the MTD to avoid potential tolerability or safety concerns. The selected dose should be high enough so that it will provide the best hope for the study product to demonstrate efficacy. The major advantage of this design is that if the test product does not deliver the efficacy as expected, a single PoC study will provide sufficient information to stop the entire development program. Thus the sponsor can allocate resources (investments and manpower) into developing other product candidates that are more promising. On the other hand, the disadvantage of such a design is that if the PoC demonstrates the efficacy of the study drug, there is quite limited information to carry the candidate product into the next study.

Usually after the concept is proven for a test product, the natural next step in the Phase II clinical development is to run a dose-ranging study in order to evaluate the product efficacy and safety at various doses. A dose-ranging study can be viewed as the first dose-response study and it typically includes a high dose, a placebo control, and a few doses in between. The high dose would be a dose that is close to MTD, or a dose not much different from the one used in the PoC study. However, the choice of the low dose or other doses between placebo and high dose can be a difficult challenge. Some researchers (e.g., Wong et al. 1996) suggest an equal-dose spacing while some suggest log dose spacing and Fibonacci series. In addition, the modified Fibonacci series approach can also be considered (Penel and Kramar 2012), along with the approach suggested by Quinlan and Krams (2006). An equal-dose spacing can be achieved by dividing the high dose by the number of test doses; for example, if the high dose is 90 mg, and there are three doses, then the medium dose would be 60 mg and the low dose would be 30 mg. The log-dose spacing allows a logarithmic relationship; for instance, if the high dose is 90 mg, then the medium dose would be 30 mg and the low dose would be 10 mg.

Hamlett et al. (2002) recommend a binary-dose spacing for dose allocation. Basically, this proposal starts with MTD. Without loss of generality, the MTD can be denoted as 1 (100%). If the dose-ranging study uses two test doses and placebo, then the high dose is allocated to be between 0.5 and 1, and the low dose is allocated between 0 and 0.5. One example is 25 and 75% of MTD.

Now suppose the design is for three test doses and placebo. Then the high dose would stay between 0.5 and 1, the medium dose would be between 0.25 and 0.5, and the low dose would be between 0 and 0.25. One example is that low dose can be at 12.5% of MTD, medium dose at 37.5% of MTD, and the high dose at 75% of MTD. When more doses are considered in this dose-ranging design, the spaces are further divided to the low end and every time the reference point is divided by 2. Hence the name binary-dose spacing.

Considerations in designing PoC and dose-ranging studies will be elaborated upon in the next few chapters.

3.9 Treatment-by-Factor Interaction

Subgroup analyses are very popular after clinical study results are available, and many of them are performed on a post hoc basis. However, from a study design point of view, most of the clinical trials are designed to study the entire subject population recruited using the inclusion/exclusion criteria. The primary objective of a given clinical trial is rarely to test a statistical hypothesis on any subgroup of the selected subject population. Accordingly, it is typical to design trials primarily for the understanding and assessment of the entire study population and only consider subgroup analyses as secondary or exploratory. After the blind is broken and clinical results are ready for data analysis, post hoc subgroup analyses are often performed.

At the study design stage, if particular subgroups are thought of as important, or specific factors could be potentially interacting with the study treatment effect, then stratified randomization might be employed. In a stratified randomization, patients are randomized based on their baseline characteristics; for instance, if sex is used as a stratification factor, then male patients are randomized to each treatment group; separately from female patients. This way it is expected that treatment allocations are balanced within each sex. This will reduce the risk that most of male patients are randomized to one treatment group, while most female patients are randomized to the other treatment group. Under such a randomization, the stratification factor(s) can be included as a discrete covariate and possibly as a treatment-by-stratification factor in the primary analytical model. When a stratification factor is included in the main model, statisticians tend to ask the question about whether there is a treatment-by-factor, or treatment-by-covariate, interaction. Does the treatment effect depend on levels of a stratification factor?

Before the early to mid-1990s, traditional clinical trials are designed with block randomization having test products pre-packaged in blocks and shipped to investigator sites (or centers). Hence investigator sites or centers became a natural stratification factor, and it used to be typical that center is included in the primary model for data analysis. The International Conference on Harmonization (ICH) formulated in the early 1990s and the ICH guidelines were also drafted around that time. The most important statistical ICH guidance is ICH-E9, which was first available in 1998, at the time when center was still a popular stratification factor. Therefore, "center" was used as an example of stratification factor in the ICH-E9 discussions. E9 made recommendations about how to pre-specify a primary analysis model using center as an implicit stratification factor and discussed the use of treatment-by-center interaction.

Note that newer ways of performing randomization such as minimization or interactive voice recognition system (IVRS) were also popularized in the early

1990s. Minimization is a method for randomization which allows more stratification factors or more levels of a stratification factor to be included in a single clinical trial design. IVRS offers flexibility of randomizing patients and medications. With these newer technologies and the fact that more centers, countries, or regions are included in a single clinical trial, center effect became less popular as a stratification factor. In many of the recent Phase III trials, center is not considered as a stratification factor, and center effect is no longer included in the primary model as a covariate. Under today's environment, it helps to read and understand ICH-E9 with some of these historical perspectives. In other words, in the attempt of understanding the ICH-E9 recommendations about a "center effect," we can interpret it as a "stratification factor effect".

On this basis, the implications of the following paragraphs from this statistical guidance may be expanded to a more general setting of stratification factors. According to ICH-E9 (Sect. 3.2),

> The statistical model to be adopted for the estimation and testing of treatment effects should be described in the protocol. The main treatment effect may be investigated first using a model that allows for center differences, but does not include a term for treatment-by-center interaction [author explanation—this means that the primary analytical model include center as a main effect, but not to include the interaction term]. If the treatment effect is homogeneous across centers, the routine inclusion of interaction terms in the model reduces the efficiency of the test for the main effects. In the presence of true heterogeneity of treatment effects, the interpretation of the main treatment effect is controversial.
>
> In some trials, for example, some large mortality trials with very few subjects per center, there may be no reason to expect the centers to have any influence on the primary or secondary variables because they are unlikely to represent influences of clinical importance. In other trials, it may be recognized from the start that the limited numbers of subjects per center will make it impracticable to include the center effects in the statistical model. In these cases, it is not considered appropriate to include a term for center in the model, and it is not necessary to stratify the randomization by center in this situation.
>
> If positive treatment effects are found in a trial with appreciable numbers of subjects per center, there should generally be an exploration of the heterogeneity of treatment effects across centers, as this may affect the generalizability of the conclusions. Marked heterogeneity may be identified by graphical display of the results of individual centers or by analytical methods, such as a significance test of the treatment-by-center interaction. When using such a statistical significance test, it is important to recognize that this generally has low power in a trial designed to detect the main effect of treatment.
>
> If heterogeneity of treatment effects is found, this should be interpreted with care, and vigorous attempts should be made to find an explanation in terms of other features of trial management or subject characteristics. Such an explanation will usually suggest appropriate further analysis and interpretation. In the absence of an explanation, heterogeneity of treatment effect, as evidenced, for example, by marked quantitative interactions (see Glossary) implies that alternative estimates of the treatment effect, giving different weights to the centers, may be needed to substantiate the robustness of the estimates of treatment effect. It is even more important to understand the basis of any heterogeneity characterized by marked qualitative interactions (see Glossary), and failure to find an explanation may necessitate further clinical trials before the treatment effect can be reliably predicted. (ICH E9 1998)

The previous paragraphs from ICH-E9 statistical guidance suggest that the treatment-by-center interaction is exploratory. If this rationale is generalized from center to stratification factors, then the nature of treatment-by-factor interaction is also exploratory in nature.

In most protocols or statistical analysis plans, the description of the primary analysis model may include the main effects of treatment, stratification factor, and baseline covariate of the outcome (dependent) variable as a continuous covariate, allowing for treatment-by-factor interaction to be explored. There could be at least three reasons to consider this interaction as exploratory in the primary analytical model:

1. The primary objective is to study the entire subject population, not any particular subgroups,
2. A model selection process could potentially inflate the study-wise Type I error, and
3. There is no pre-specified alpha for testing the treatment-by-factor interaction.

The first reason has been clarified in the ICH-E9. In other words, at the study design stage, the expectation is that the test product be superior to the placebo control across all subgroups. Again, the objective, or the designed feature, is not to "confirm" any interaction or any treatment benefit to a particular subgroup. Any treatment-by-subgroup interaction or subgroup treatment effect can be considered only as an "observed feature" (as opposed to a "designed feature"). Nevertheless, if it is not known at the design stage whether any subgroup may benefit more from this test treatment, the treatment effect in subgroups or with interactions can be explored. In case it is already known that the study treatment benefits a particular subgroup and then Phase III clinical trials may be designed to only confirm treatment efficacy for this well-defined subgroup. For example, if the team believes that the study drug only benefits female patients, then the clinical study should be designed to only recruit females. Here protocol inclusion//exclusion criteria would reflect the selection of those subjects of interests. On this basis, at the design stage, there is no need of a treatment-by-subgroup interaction in the primary statistical analysis model.

From the alpha protection perspective, any model selection process could potentially inflate alpha. Suppose that two models are considered as candidates for the primary analysis. After data read out, one model shows a statistical significance in the primary treatment comparison whereas the other fails to do that. A question arises of which result should be used to make the decision from this given clinical trial. Therefore, a recommended process of including a treatment-by-subgroup interaction term in the initial model and, then after it is decided that the interaction is not significant, dropping this term from the model (in favor of reanalyzing the same data without such an interaction term) could potentially inflate alpha in practice. This approach is not applicable in designing clinical trials where the primary objective is to demonstrate clinical efficacy across all participating subjects. Hence the second point from above indicates that, in practice, model selection is not appropriate under this setting.

The third reason is about pre-specified alpha in regard to the interpretation of a "treatment-by-factor" interaction which is a natural extension from the previous point. If an interaction is included in the primary model, what *p*-value offers a "significant interaction"? Dose an observed *p*-value of 0.06 imply that there is no treatment-by-subgroup interaction? In case an alpha is allocated to test for such an interaction, how should this alpha be used? Should the alpha for the interaction test be split and use the rest to test for the main treatment effect? Note that the aforementioned paragraphs describe the justification why the interaction term should not be included in the primary analytical model. If readers are interested in the exploration of interaction, a separate discussion will be necessary, which is beyond the scope of this book.

All of the above discussions point to a simple understanding that the treatment-by-subgroup interaction, by nature, is an exploratory phenomenon in a trial; it is not likely to be confirmatory. This aspect is part of an observed feature in any given clinical study, as the interaction test does not appear to be a designed feature. On this basis, the exploration of a treatment-by-subgroup interaction can be achieved by estimation, instead of hypothesis testing.

In other words, when the interest is to learn about a particular interaction of interest, a point estimate and a confidence interval for this interaction provides more information than a simple *p*-value. For example, in a two-treatment clinical trial that is designed to compare the test product against a placebo control, if gender is considered as a stratification factor, then the primary analysis model for this study design may include treatment, and gender as main effects (and frequently the baseline covariate of the outcome) in the main model. Then either in the protocol or in the statistical analysis plan, a statement could be added such as "treatment-by-gender interaction will be explored."

After the data read out, a treatment-by-gender interaction can be estimated based on the clinical question of interest: Is there a difference in treatment benefits for those male patients as compared to those female patients? In order to answer this question, the interaction can be specified as $(\mu_{test,M} - \mu_{pbo,M}) - (\mu_{test,F} - \mu_{pbo,F})$. A point estimate of such a quantity can be interpreted as the treatment benefit for males subtracted by the treatment benefit for females. Then appropriate confidence intervals can be constructed for such a parameter. The more informative presentation could be a sequence of confidence intervals such as 80% confidence interval (CI), 90% CI, 95% CI, and 99% CI. This analysis provides not only the direction of the interaction (more benefit to males? or more to females?), it also quantifies the interaction. For exploratory or learning purposes, a *p*-value does not offer sufficient information to the readers.

Another worthwhile point is that the mixed model with repeated measures (MMRM) has become a popular model in analyzing longitudinal clinical data (Mallinckrodt et al. 2008). When this model is used as the primary model, another term—visit—is included as a categorical covariate in the primary analysis model. Accordingly, a treatment-by-visit interaction, as well as baseline covariate (of outcome)-by-visit interaction term, may be added to this model. The interpretations of these terms are very different from the interpretation of the

treatment-by-subgroup interaction. Note that if the model does not include a treatment-by-visit interaction, that implies a strong assumption that the treatment differences are constant across visit. If the team is not willing to make this strong assumption, then a treatment-by-visit interaction is necessary in the longitudinal model. This, of course, is very different from the treatment by subgroup interaction.

First, when a treatment-by-visit interaction is added to this model, what does this interaction mean? Suppose a very simple model includes only the main effect of treatment and main effect of visit. Then the estimate of treatment provides information regarding the treatment responses to each treatment group and treatment differences between groups of interest, across all visits. The estimate of a visit effect offers estimates at each visit across all treatment groups.

Still, in an MMRM analysis, the primary interest is typically a treatment benefit at the primary time point of interest. Such a simple model assumes treatment effects are constant over time; that is, treatment benefit at the first visit post-baseline is the same as that what follows later at the primary time point of interest.

Suppose a trial is designed with monthly visit and the primary time point is at month 6. Under such a study design, the assumption would be that the treatment benefit of test product as compared to the placebo control improves over time, and that up to month 6, such a benefit would be clinically meaningful. In other words, it is not expected that the treatment benefit at month 1 is strong enough and, in addition, that patients will need to be exposed to the test product over 6 months in order to experience the full benefit of such a product.

A treatment-by-visit interaction allows the treatment benefits to change over time (visits). Hence by adding this interaction term, the model users are relaxing an assumption of a constant treatment effect over time. Doing so is more realistic and less restrictive.

A baseline-by-visit interaction implies that the impact of baseline to responses at different visits could be different. Again, in a six-month trial with only main fixed effects of treatment, visit and baseline covariate of outcome, without any interaction term, the implication is that the baseline effect to patient responses at the one month visit is the same as that to the six months after randomization. In common analytical models using baseline (of outcome) as a covariate, the assumption is that baseline values is correlated to corresponding values in post-baselines outcome at the primary time point of analysis.

On this basis, it can be understood that the treatment-by-visit interaction, as well as the baseline-by-visit interaction, is part of the designed feature associated with a longitudinal study. Including these terms in the MMRM for the primary analysis is scientifically justified. It is also noteworthy that there is no model selection activities in such an MMRM as the interaction terms remain in the primary model regardless of the corresponding observed p-values for these terms.

The understanding of treatment-by-subgroup interaction is exploratory in nature. It is generally not recommended to include such a term in the primary analytical model. However, in a secondary analytical model, both the interaction analysis and the subgroup analysis are viewed as exploratory. Estimation procedures are appropriately applied to learn these results. Point estimates, along with confidence

intervals, can be constructed on the parameter of interests. For subgroup analyses, descriptive statistics can be very useful. For treatment-by-subgroup interaction, whenever there is a particular single degree of freedom interaction needs to be studied, appropriate confidence intervals can be established. For example if there is only two treatment groups and two subgroups, let μ_{ij} represent the subgroup mean of the *i*th treatment and *j*th subgroup. Then the interaction can be expressed as $\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$, after which the point estimate and interval estimate can be obtained for this contrast.

3.10 Evaluation of Product Safety

Before the story of lady tasting tea and the concept of randomized controlled clinical trial, it was very difficult to objectively demonstrate product efficacy. In fact, a medicinal product would be approved, from a statistical point of view, because "the probability that the product being a placebo is controlled under alpha." If the demonstration of product efficacy is very difficult, then the proof of product safety is almost impossible. Therefore, most of the approved products have "their benefits outweigh their risks" mainly because to scientifically prove product efficacy and product safety are extremely difficult.

Product safety does not tend to be a comparative concept. Hence "the safety profile of the test product is no different from that of a placebo" may not be sufficient to demonstrate the safety of a test product. Although safety profiles are compared between the test intervention and the control intervention in a clinical trial, "relative safety" itself is not enough to establish the safety of the test product. Product safety is not necessarily a population-wide phenomenon—in some situations, a very severe individual event may cause the development of a particular product candidate to be stopped. For example, in the development of a candidate product indicated for a certain type of chronic pain, if treatment-related Stevens-Johnson Syndrome is observed from a few patients, then the entire development program may have to be stopped.

Finally, patient demographic data and product safety is an observed feature, not a designed feature. For a clinical trial, primary and secondary endpoints are generally considered as efficacy variables. All other observations such as patient demographic data, baseline characteristics or safety variables are considered exploratory because they are not part of the designed feature. In other words, these results are observed phenomena but not designed phenomena. All of these discussions imply that the thinking and the understanding of patient demographic data and product safety are different from that of product efficacy. Hence safety and demographic data do not follow a statistical hypothesis testing framework; it is exploratory in nature.

Product safety is one of the most important characteristics to be followed and observed for every medicinal product. In fact, during the entire product discovery and development process, safety is being studied very closely at every stage. One of

the critical non-clinical development activities is product toxicity. As discussed in Chap. 1, animal toxicity studies are conducted for various species of animals at multiple doses with different duration. The need to understand product safety continues during clinical Phase I, II, and III development; even after the product is approved and available for general patient population use, Phase IV studies and pharmacovigilance studies are still designed to follow product safety.

Again, because product efficacy is evaluated in the form of statistical hypothesis testing, it follows a confirmatory way of thinking. In contrast, patient demographic data and product safety can be thought of as exploratory. The main point of this chapter is that in designing clinical trials, especially Phase II and Phase III trials, the confirmatory objectives (e.g., treatment efficacy of the candidate product) are addressed first and only after the product efficacy is confirmed, then the safety profile of the test product can be explored in a greater degree.

3.11 Every Clinical Trial Can Be Considered as Both Confirmatory and Exploratory

It is generally believed that Phase III clinical trials are designed for regulatory decision making; therefore, these studies are thought of as confirmatory trials. Accordingly, Phase I and II clinical trials are designed to explore various characteristics of the product candidate. It would be natural to think of these studies as exploratory ones. Note that, in certain situations like life-threatening diseases (e.g., oncology), there may be no Phase III for drug approval: a drug may get approved at Phase III. Or Phase II may be skipped and the drug accelerated to Phase III. In fact, After Phase III studies are completed and data are summarized, analyzed and reported, researchers can still learn about the medicinal product under study. Thus, in a certain extent, Phase III studies are exploratory In many cases, new hypotheses could be generated based on Phase III clinical trial results, and additional studies are designed to help further understanding the product. Similarly, Phase I and Phase II studies are usually used for internal decision-making within the sponsor and, consequently, Phase I and II clinical trials possess the confirmatory features too. Therefore, every individual clinical trial can be looked at as both exploratory and confirmatory depending on the purpose at hand.

Because the entire scientific process involves a evolution of learning, it is easy to understand why researchers learn from every clinical trial and thus that each trial is an exploratory trial. It could be difficult to consider every clinical trial also as confirmatory.

Sample size calculation usually is based on a statistical hypothesis. Sometimes the sample size of a given study is obtained from confidence intervals. In general, when there is a hypothesis that is used to calculate the sample size, this hypothesis reflects the confirmatory feature of the study. On the other hand, if sample size is calculated from a confidence interval, then it will be of interest to understand the

primary objective of this study—if the objective is exploratory, then estimation can be applied for data analysis. However, if the objective is confirmatory, then the confidence interval can be translated into a hypothesis, and the traditional hypothesis testing methods can be used for data analysis.

Recall from Chap. 1 that the most efficient strategy in medicinal product development would be a "design to stop" concept. The confirmatory objective of each study could be thought of as a Go/NoGo decision. On this basis, it might be easier to think that the confirmatory objective of each study is a Go/NoGo decision and, only after the "Go" decision is confirmed, other efficacy endpoints, safety, interaction, or various doses can be explored from this given study.

Following this thread of reasoning, we can consider every clinical trial design as "confirmatory" first implying a Go/NoGo decision is made about the candidate product. If the decision is NoGo, it is a very efficient way to weed out a weak test product at the earliest stage. On the other hand, if a clear "Go" decision is made about this candidate, then the sponsor could assume this test product is good enough to move to the next step of development. Meanwhile, various characteristics about this candidate can be explored. Hence for those studies designed to answer a confirmatory question, the design can be viewed as a process that is first "confirm" and then "explore."

3.12 Conclusion

This chapter discusses the paradigm of drafting a clinical development plan and designing individual Phase II or III clinical trials. Traditionally, in general, the scientific learning process begins with exploration and then confirmation, followed by more exploration and then more confirmation, and so on. But in the clinical development program of a new medicinal product, this thinking process is reversed. A clinical plan needs to start with a draft label and work backwards to consider what Phase III data are necessary to support the language in the drug label. Considerations are given to how Phase II studies can deliver the data to help with Phase III design. In order for Phase II studies to deliver the data that will help Phase III design, certain information needs to be available from Phase I. In designing each individual Phase II or III clinical trial, the statistician and the team can proceed with the thinking process that starts with confirming the clinical question of interest. After this step is accomplished, the next step is to explore additional findings.

However, establishing the confirmatory step is not the only important aspect of clinical development. Studies also need to be designed to explore other characteristics of the test drug, which is especially important in the Phase II proof of concept and dose-finding stage. A motivational example is introduced to demonstrate why it is important to consider this stepwise paradigm in designing Phase II/III clinical studies. Finally, a few additional examples are discussed to show how

this thinking process can be applied critically and judiciously in practical study design situations.

References

Dmitrienko, A., Wiens, B., & Westfall, P. (2006). Clinical trials. *Journal of Biopharmaceutical Statistics, 16*(5), 745–755.

Fricke, J., Varkalis, J., Zwillich, S., Adler, R., Forester, E., Recker, D. P., et al. (2002). Valdecoxib is more efficacious than rofecoxib in relieving pain associated with oral surgery. *American Journal of Therapeutics, 9*(2), 89–97.

Hamlett, A., Ting, N., Hanumara, C., & Finman, J. S. (2002). Dose spacing in early dose response clinical trial designs. *Drug Information Journal, 36*(4), 855–864.

ICH E9 Statistical Principles for Clinical Trials. (1998). International conference on harmonization —Harmonized Tripartite Guideline.

Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y., & Mancuso, J. P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal, 42,* 303–319.

Penel, N., & Kramar, A. (2012). What does a modified-Fibonacci dose-escalation actually correspond to? *BMC Medical Research Methodology, BMC series, 103*. doi:10.1186/1471-2288-12-103

Pinheiro, J. C., Bretz, F., & Branson, M. (2003). Analysis of dose-response studies—Modeling approaches. In *Dose finding in drug development* (pp. 146–171). New York: Springer.

Quinlan, J. A., & Krams, M. (2006). Implementing adaptive designs: Logistical and operational consideration. *Drug Information Journal, 40*, 437–444.

Ruberg, S. J. (1995). Dose response studies II. Analysis and interpretation. *Journal of Biopharmaceutical Statistics, 5*(1), 15–42.

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

Ting, N. (2003). Introduction and new drug development. In *Dose finding in drug development* (pp. 1–17). New York: Springer.

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

Wong, W. K., & Lachenbruch, P. A. (1996). Tutorial in biostatistics: Designing studies for dose response. *Statistics in Medicine, 15,* 343–359.

Chapter 4
Design a Proof of Concept Trial

4.1 Introduction

In typical clinical development programs, Phase II is the first time for the candidate product (compound or biologic) to be tested in patients with the target disease under study. For medicinal products discovered and developed to treat chronic diseases, most of the Phase I clinical trials recruit normal healthy volunteers and, as a result, the disease improvement cannot be observed in Phase I. Under this situation, therefore, study results cannot be used to assess product efficacy. In Phase I oncology trials, although cancer patients are recruited to help study the maximal tolerable dose (MTD), these patients may or may not be with the tumor type for which the compound is developed for. Also, the sample sizes used for Phase I oncology trials are not large enough for study of product efficacy. Therefore, in most of product development programs, Phase II is the first time the product efficacy can be observed and evaluated. In other words, before the first Phase II clinical trial, the sponsor does not have an opportunity to study whether or not the test product, or compound, works in human patients.

Under this circumstance, it is especially important that the product efficacy be established from the first Phase II clinical trial. The most commonly used study design for a first Phase II clinical trial is a proof of concept (PoC) study. A PoC typically includes two treatment groups: a placebo group and a test product treatment group (for PoC designed with more than two groups, please refer to Chap. 6). In order to offer the best opportunity for the test product to demonstrate its efficacy, the dose of test product used in a PoC tends to be the highest possible dose. And this dose is usually the MTD, or a dose that is slightly below MTD. This design allows the project team to make a Go/NoGo decision on the test product after study results read out.

If the results fail to demonstrate a clear efficacy, then a No-Go decision can be made for this candidate product, and further development can be stopped. The sponsor does not have to make more investment on it. Budget and resources can be

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_4

allocated to develop other candidates that show more promise for success. This approach can be also considered as a "sledgehammer" approach, because such a simple design helps simplify the conclusion to be a "make it or break it" result for the product under development; that is, a clear Go/NoGo decision should be made based on analysis of data obtained from the PoC study.

Another type of study is known as a "Proof of Mechanism" (PoM) study. In the laboratories in which basic research and discovery are performed, it is important that a good understanding of the "mechanism of action" be available for the compound that is being discovered. For example, before the beta blockers were developed for the treatment of high systolic blood pressure (SBP), the biologic theory was that the beta receptors in heart muscles may increase SBP as a response to stress stimulants. Given this understanding, the lab scientists searched for a chemical structure that could bind to the beta receptors and inhibits their activities. Such a chemical compound could be considered as a "beta blocker" because it blocks the activities of beta receptors in heart muscles.

As discussed earlier, product efficacy cannot be established until Phase II. However, in the case of developing a beta blocker, there could be opportunities to observe whether the mechanism of action can be demonstrated in a Phase I clinical trial. This query can be achieved by observing the activities of beta receptors in healthy normal volunteers. In other words, although healthy subjects with normal SBP are recruited in Phase I, activities of beta receptors from these subjects can still be measured. By assessing beta receptor activities from these subjects, a proof of mechanism objective could potentially be achieved. Often times the measurement that helps establishing PoM is denoted as a biomarker. Generally speaking, PoM is not a well-defined, widely accepted term that applies to all candidate products. Only under certain situations where the marker can be easily identified, and clearly measured, a PoM study design would be contemplated.

Traditionally, cancer treatments are thought of as "chemotherapy," which is grounded in the thinking of eliminating cancer cells in the human body. In fighting cancer, the first step is to remove the tumor (either by surgery or by radiation), and then to follow it by chemotherapy to destroy the remaining tumor cells which were not removed by surgical or radiation procedures. Most of those drugs which were developed to destroy cancer cells could, unfortunately, also be harmful to normal tissues. These drugs belong to a class of "cytotoxic" medications, implying that these drugs are toxic to normal cells too. In the past few decades, a newer class of anti-cancer agents has been developed and marketed. This new class of medications is considered as "cytostatic." The concept of developing cytostatic drugs is not to kill tumor cells or normal cells; it is to allow tumor cells in human body, by merely suppress their growth and their activities. One example could be the discovery and development of vascular endothelial growth factor (VEGF) inhibitors. Certain tumor cells attract VEGF so that blood vessels are over grown to supply blood to tumor tissues, instead of normal tissues. VEGF inhibitors are developed to slow down the growth of these blood vessels and hence tumor tissues do not grow as fast as normal tissues. From this point of view, a VEGF inhibitor is considered as cytostatic because it is not harmful to normal cells.

In the development of cancer treatments, because most of the compounds could be cytotoxic, it is not ethical to expose healthy normal volunteers to this class of test products. Hence Phase I cancer clinical trials recruit cancer patients as study subjects. Consequently, placebo is not appropriate to be used as the control agent. In typical product development programs for treating cancer, Phase I clinical trials are designed to explore MTD. These trials recruit cancer patients, and escalate test product doses from low to high, without a placebo control.

One of the popular designs is the 3 + 3 design (Ivanova 2006), which allows cancer patients to be exposed to escalating doses from low to high. MTD is usually achieved or exceeded when one or several dose limiting toxicities (DLT) are observed. One reason that cancer patients, instead of healthy normal volunteers, are recruited for Phase I cancer trials is that for a life-threatening disease like cancer, patients are willing to tolerate more severe adverse events. For example, among the typical adverse events associated with chemotherapy include nausea, vomiting, and hair loss. For cancer patients, they are willing to tolerate these events because the condition they are facing is life-threatening. However, for another indication such as pain or high cholesterol, subjects may not be willing to tolerate these types of events. Consequently in oncology Phase I trials, it is typical to recruit cancer patients, instead of healthy normal volunteers.

Therefore, Phase I clinical development programs for compounds treating cancer is very different from compounds developed in treating other, non-life-threatening conditions. In oncology, the primary objective of Phase I development is to estimate MTD, which is accomplished by observing DLT's from cancer patients. For clinical development programs of products treating non-life-threatening conditions, the objectives for Phase I clinical trials are estimating MTD from normal healthy volunteers and to characterize the pharmacokinetics properties of the candidate drug under development. However, after Phase I, the first Phase II study is usually a Proof of Concept (PoC) study regardless whether the compound is developed to treat a life-threatening disease or not. At this stage, the organization (usually the sponsor) that develops this compound is eager to learn whether such a candidate product can really deliver efficacy in treating patients with the underlying condition. Running a PoC study is a very sensible first step in Phase II.

4.2 Proof of Concept Trials

4.2.1 Impact of PoC Decisions

Most PoC studies tend to have two treatment groups: one placebo control group, and another treatment group with a high dose of the candidate product. In recent years, some PoC studies are designed with more than two treatment groups. These studies will be discussed in Chap. 6. Generally speaking, the PoC designs are based on two key assumptions: (1) the MTD estimated from Phase I is appropriate and (2) efficacy dose-response relationship is monotonic.

Under these two assumptions, the highest possible dose (either MTD or a dose slightly below MTD) of the test product offers the best opportunity to demonstrate efficacy response. In other words, if this dose does not show efficacy response, then there is no need to invest any more resources to develop this test product. Hence, only after the concept is proven, extended development activities can be funded and resources can be allocated to further develop the compound under investigation.

A positive PoC result triggers many important development activities. A long-term toxicity study is one of those activities. Drug formulation and drug supply also depend upon a positive PoC. At early Phase II, there is usually sufficient drug supply to handle the first few dose-ranging studies. If the PoC result is positive, there will be more study medication needed for Phase III clinical trials, and large amount of raw materials need to be ordered or to be manufactured. Moreover, depending on the dosage forms that may be necessary for additional Phase II trials and Phase III trials, various formulations of the candidate product will have to be prepared. On the other hand, if the concept is not proven, there is no need to stock up a huge pile of raw materials for the making of large quantities of the study drug or study biologic.

Furthermore, additional bioavailability or bioequivalence (or both) studies may be necessary. In some situations, additional drug-drug interaction studies may also be needed. Therefore, a clear Go/NoGo decision is critical from the PoC study results. When the study results are not clear (inconclusive), many of the important development activities will be delayed. Challenges of inconclusive PoC results are covered in Chap. 7.

A typical early phase clinical trial tends to be designed with relatively short duration. Many Phase I trials are single-dose trials, which means only one single dose is given to the trial participants. Some other trials may be designed with treatment duration of one week or two weeks. For products developed to treat non-life-threatening conditions, the limiting factor of time length comes from the pre-clinical (animal) toxicity studies. If results from two-week animal toxicity studies indicate the test drug is negative with respect to toxic findings, then a clinical trial can be designed with two-week duration so that trial participants are not subject to excessive test product exposure.

According to the clinical development plan for a study medication, if the Phase III study needs to be of one-year duration, then (before the Phase III study is started) a one-year toxicity study on animals should be completed and with no safety findings. Usually such a one-year toxicity study will not be initiated unless the concept is proven from early Phase II. In practice, if the Go/NoGo decision is delayed because of inconclusiveness, then all these operations could be postponed and, eventually, this problem of inconclusiveness may delay the entire product development program.

4.2.2 How to Communicate Risks Associated with a PoC Study

Usually a statistician is involved in a study design when there is a need for sample size calculation. Sample size estimation is one of many essential contributions provided by a clinical trial statistician. An experienced statistician will begin by understanding the primary endpoint and by probing its clinically meaningful difference, delta (δ). Sometimes this quantity is readily available. However, in many cases, this could be the beginning of long discussions.

After the study completes, clinical data are analyzed and results are presented to the project team. The project team members and upper management evaluate these results and make a decision. A Go/NoGo decision is easy if both statistical significance is observed and treatment difference exceeds δ, where δ is the postulated clinically meaningful difference. These results indicate a clear "Go" decision. On the other hand, when the results are not statistically significant, and the observed treatment difference is too small, then the team may make, often times reluctantly, a NoGo decision. However, if the observed treatment difference is not too small, but failed to reach statistical significance, then the team is in a difficult position and not clear as to whether a "Go" or a "No-Go" decision could be made with confidence. Under this circumstance, a situation of inconclusiveness may take place.

In many cases, when the observed *p*-value is close to statistical significance, but not strictly statistically significant, the team tends to decide to move forward. However, future development could indicate this "Go" decision was a mistake because the observed treatment difference was not large enough. These difficulties occurred because, at the time of designing PoC, a delta which is larger than the true desired treatment difference was used for various reasons. One typical case could be that such a δ was selected based on the available sample size; that is, from the sample size and the given α and β, a treatment difference can be calculated. A larger δ was chosen so that such a PoC study was feasible under the given budget constraints. However, after the PoC results are ready, although the observed treatment difference was actually smaller than the hypothesized δ, the project team may still make a "Go" decision simply because the *p*-value was either statistically significant or close to be significant.

On the other hand, there were situations where the decision was NoGo even when a statistical significance is observed, yet the clinically meaningful difference δ was not achieved. One example can be found when the PoC endpoint is different from the Phase III endpoint. When this is the case, a lowest δ was selected for PoC and the decision rule was that, if the PoC results cannot deliver such a δ, then there is no hope for the test product to meet the Phase III primary endpoint.

Under certain situations, the PoC study may have to be repeated so that a clear decision can be made. Inconclusiveness can be thought of as the time period between the study results read out and a definitive "Go" or NoGo decision is made. During this period, the team could consider a "Go" decision, even when the observed *p* is greater than α. On the other hand, if the observed *p* is less than α, but

the observed treatment difference is less than δ, but a NoGo decision was made incorrectly (β could be inflated). In either case, the risk of making a wrong decision is greater than the pre-specified error rate.

Based on the above discussion, it is clear that when there is an observed statistical significance, and the decision is Go, then the risk of making a wrong decision is preserved under the pre-specified level alpha (α). Also, when statistical significance is not observed, and the decision is NoGo, then the risk of making a wrong decision is also preserved under the pre-specified level of beta (β). However, if a Go decision is made when there is no statistical significance, then the risk of making a wrong "Go" decision is increased to be more than the pre-specified risk level. At the other end, when there is a statistical significance and a NoGo decision is made, then the risk of making a wrong NoGo decision is increased to be more than the pre-specified β. Of course, these decisions are usually made from long discussions among team members after looking at all clinical data on hand—primary and secondary endpoints, multiple doses if more than one dose was used, safety data, sometimes pharmacokinetic (PK) or outcomes research data, and so on.

The main issue we hope to address in this chapter is during the time between the PoC data are ready, and the actual decision is made, there is a period of time that the PoC study was considered as "inconclusive." This time period is difficult because there is always tight timelines for the entire project. Will there be a need for a long-term toxicity study? when should it start? What amount of drug supply will be necessary? At what time frame (Phase II requirements and Phase III requirements)? What dosages to be formulated? Will there be additional PK studies such as drug-drug-interaction or food effect studies to be designed? At this time, a Go/NoGo decision is very critical. Yet the data are not always clear. Regardless of all these discussions and various perspectives, in the end, there will have to be a decision—to go or not to go. This problem becomes much more difficult when the PoC study needs to be repeated.

Although the design of a PoC study is considered a sledgehammer approach, in practice there could still be many difficulties and challenges in decision-making after the study is completed. Below are some important points need to be considered before designing a PoC clinical trial. It is critical to keep in mind that in the design of such a study, there should be sufficient discussions among team members with the objective of minimizing the risk of potential inconclusiveness.

4.3 The Primary Endpoint in a PoC Design

As mentioned previously, the duration of treatment for a clinical trial is limited by the length of time covered by toxicology studies from the non-clinical development stage. At early Phase II, trials are usually designed with duration of a few weeks or a few months depending on the therapeutic area, and characteristics associated with

the test product being studied. Accordingly, the primary endpoint of Phase II trials tends to be a clinical efficacy variable, usually measured as relatively short-term patient response to the interventions.

In certain trials, the primary endpoint could be a time-to-event variable (for example, oncology trials). However, the primary efficacy endpoints for most of Phase II clinical studies are measured as a continuous variable or a categorical variable. For products developed to treat chronic diseases, the primary endpoint is typically a change from baseline. For example, if the indication is hypertension (high blood pressure), then the primary endpoint would be the change in systolic blood pressure (SBP) or diastolic blood pressure (DBP) from baseline to the primary time point (e.g., four weeks after double-blind treatment). For a product developed to treat depression, the primary endpoint could be the change in Hamilton Depression Rating Scale (HAM-D) from baseline to the primary time point.

In analysis of a continuous variable such as changes in SBP, DBP or HAM-D, the analysis of covariance model is typically employed. In a development program for a test product to treat rheumatoid arthritis (RA), a Phase II primary endpoint could be the number or proportion of ACR20 responders. ACR20 is proposed by the American College of Rheumatology (ACR) based on a summary of seven clinical measurements—if an RA patient experiences at least a 20% improvement using the ACR algorithm of these seven measures, that patient is considered as an ACR20 responder. Otherwise the patient is considered as a non-responder to ACR20.

If the primary endpoint is a binary variable such as proportion of ACR20 responders, then a two-by-two table (two treatment groups and two responding status—yes or no) can be formulated from the study results. There are many different ways of analyzing two-by-two tables—normal approximation of binary responses, difference in proportions of responders, ratio of proportions, or Fisher's exact tests.

4.4 MTD Could Be Under Estimated or Over Estimated

One of the most important deliverables of Phase I clinical trials is an estimate of maximal tolerable dose (MTD). The understanding of MTD is that a subject may experience adverse events or tolerability issues if he or she receives doses of the study product that is above MTD. In clinical development programs for products treating non-life-threatening diseases, Phase I experimental units are normal healthy volunteers. Unfortunately, in many practical cases, the MTD obtained from Phase I turns out to be either higher or lower than the true MTD, which gets subsequently determined at a later stage of clinical development. One question could be that, if the MTD was found based on healthy normal volunteers, can it be translated to patients appropriately?

If MTD gets overestimated, then patients recruited in Phase II clinical trials could be exposed to excessive toxic doses and a large number of adverse events could be observed. This miscalculation may lead to a pre-mature stopping development of a potentially good candidate or a slowdown in the development process.

On the other hand, underestimation of MTD creates a more difficult problem in product development. In this case, the dose used at PoC may demonstrate a weak signal as compared against placebo. Phase II clinical data may indicate that a dose higher than the perceived MTD could be relatively safe, and deliver more efficacy. When this is the case, multiple studies can be designed to help the doses to creep higher and higher until a true MTD can be determined. Consequently, underestimation of MTD from Phase I could introduce major inefficiencies in product development. One of the most expensive delays in clinical development programs is re-work. In the case of underestimation of MTD, time consuming re-work could potentially drag the development process backward.

One common confusion stems from the assessment of success or failure of a clinical project team is linked with the success or failure of the study product. In other words, people tend to consider the team is successful when the compound is successful, and the team fails when the compound fails. This is, in fact, a wrong perception. In many cases when a very strong team that designed and carried out very scientifically sound clinical trials, yet the study product failed. Under this circumstance, it is not fair to claim the team failed.

One of the metrics in assessing a project team is to see if a clinical trial needs to be repeated. The problem of underestimation of MTD is a good example—the project team in Phase I failed to deliver a good estimate of MTD. Another example could be that a dose ranging trial was designed without a placebo control. In this situation, if there are three doses and the high dose is more efficacious than medium dose, and the medium dose is more efficacious than the low dose. Then the study product was progressed and later people found that even the high dose does not provide better efficacy than placebo. This example indicates that the test product should have been stopped for development if the dose ranging trial included a placebo control in it. Therefore, it is critically important to recognize that whether the test product is successful or not, it should not be used as a yard stick to measure the performance of the study team.

Nevertheless, there is always a need to establish an anchor for the upper dose range to help design the next Phase II study. An estimate of MTD remains necessary, and the only way to move forward is to assume that the MTD estimated from Phase I is correct.

In certain drug development programs, when dose-escalation studies from Phase I pushed the dose up very high, but no adverse event was observed. Practically, if the dose is too high so that the pill is too big to swallow, or that there is a formulation limit, then higher doses became not feasible. When this is the case, the escalation study still needs to stop at a relatively high dose. This dose is denoted

as the maximally feasible dose (MFD). Hence in order to progress a test product to Phase II, it is important for the project team to assume either the MTD, or the MFD, is correctly estimated.

4.5 Monotonicity Assumption

4.5.1 Background

The fact that typical PoC studies are designed with the MTD of test product comparing against a placebo control is based on the monotonicity assumption. In other words, the hidden justification of choosing the highest tolerable dose for proof of concept is that it is assumed that this dose offers the best possible opportunity to show efficacy among all of the candidate doses to be selected. The question is, is it reasonable to assume a monotonic efficacy dose-response relationship?

For product safety, or toxicity, it is generally believed that the safety issues (adverse events, lab abnormalities, or others) increase as doses increase. This belief is why MTD serves as an anchor for the upper end of the dose range to be studied. The reason behind such a belief, or such an assumption, is that every medicinal product is toxic—if a product could cure a disease, or improve a particular health condition, it must have changed the biological system in human body. If it changes the system, it could also cause problems to the human body. Hence by increasing the amount of exposure, it is expected that the potential safety problems or issues will also increase—a monotonicity assumption.

In most of the disease areas, this same assumption is also applicable to product efficacy. If the medicinal product helps improve the disease or health condition, then more of such a product would help improve the condition better. For example, if a compound is developed to reduce pain, then it is expected that more of such a compound in the body could translate to more pain reduction for the patient who suffers from pain. This idea is applicable in most therapeutic areas. Therefore, unless it was proven otherwise, the generally accepted assumption is that as doses increase, efficacy responses also increase—again, the monotonicity assumption.

However, one well-known therapeutic area that does not assume monotonic efficacy dose-response relationship is the class of anti-psychotic products (for example, this phenomenon can be found in drug labels of Lexapro, Celexa, Paxil, Zoloft, Prozac or other antipsychotic drugs). Many of the candidates developed for the treatment of psychiatric disorders demonstrate a non-monotonic dose-response relationship. The phenomenon of non-monotonicity in this therapeutic area can be observed not only in animal models but also in clinical experiences. It can also be observed across many different anti-psychotic medicines with various indications (anti-depression, anti-anxiety, and other conditions) and with various mechanism of actions.

4.5.2 Strong or Weak Application of the Monotonicity Assumption

The assumption of monotonic efficacy dose-response relationship can be applied with different ways. For example, one of the popular multiple comparison-procedures (MCP) is known as the gate-keeping procedure. One of the uses of the gate-keeping procedure in dose-response studies is to sequentially test each dose, from high to low, and that each dose is tested against the placebo control with the entire alpha applying to each pairwise comparison. Under the monotonicity assumption, the highest dose is first tested against placebo. If there is no statistical significance, then the MCP stops and claim none of the doses is different from placebo. If the null hypothesis that the highest dose is not different from placebo is rejected, then declare the highest dose to be statistically significantly different from placebo, and the entire alpha is left for the testing of the second highest dose against placebo.

This testing procedure is continued until either there is lack of statistically significance (at the full level of alpha) or until all doses are compared against placebo. Under this application of the monotonicity assumption, if any of the lower doses become efficacious, while one of the higher doses fails to demonstrate statistical significance, the decision to stop further testing would prohibit the lower doses to even be tested. Hence such an application of the monotonicity assumption can be considered as a strong assumption of monotonic efficacy in the dose-response relationship.

In practice, variability could be observed among various test doses of a compound. Even if the underlying dose-response is truly monotonic, the observed treatment effect could still look non-monotonic. One statistical tool was developed to deal with this situation is called isotonic regression (Robertson et al. 1988). The main idea of isotonic regression estimates is pool adjacent violators (PAV). Based on the strong assumption of monotonicity, the isotonic regression forces the estimate of a treatment group mean obtained from a higher dose to be not lower than that of a relatively lower dose (assuming that higher scores are better). For example, in a dose-response study with placebo and four test doses, denoted as d_0, d_1, d_2, d_3 and d_4, in the order of doses; that is, d_0 represents placebo, d_1 represents the lowest dose, d_2 the next lowest, d_3 the next one, and d_4 represents the highest dose. Suppose the respective observed mean responses at each dose group are 5, 7, 9, 8, and 12, respectively. In this result, the observed response of dose d_3 is 8, while the observed response of dose d_2 is 9, which violates the monotonicity assumption. The isotonic regression estimate forces monotonicity by pooling the observed means of doses d_2 and d_3. Under PAV, the isotonic regression estimates of these two doses are 8.5 and 8.5 (the average of 9 and 8).

The above two cases—the use of gate-keeping procedure to control alpha adjustment in MCP and the use of isotonic regression to force monotonic dose-response relationship—are examples of assuming monotonicity in a strong sense. However, in the design of a PoC study, with the MTD selected as the dose to compare against placebo, this could be thought of as a weak application of this

monotonicity assumption. This occurs because, even if a dose lower than MTD demonstrates a higher level of efficacy than the MTD, as long as the MTD is significantly different from placebo, those lower doses will still have an opportunity to be evaluated in future studies.

In the discovery laboratories, if a compound or a biologic does not demonstrate a clear and strong monotonic concentration-response relationship, that candidate would not likely progress into clinical development. Therefore the assumption of a monotonic dose-response relationship in efficacy from human beings is not only intuitive and natural; it also has its scientific basis. Nonetheless, non-monotonic dose-response relationships are actually observed in clinical trials. Given these difficulties, before designing a PoC clinical study, can monotonicity be assumed? Our belief is that in a PoC design, we cannot afford not to use the monotonicity assumption. What follow is a set of rationales to support this line of reasoning.

4.5.3 Why This Assumption Is Still Useful

In the early Phase II stage, the upper dose range for exploration is bounded above by the MTD. Moving into Phase II, one of the objectives is to evaluate efficacy responses observed from patients treated with the candidate product, and the range of possible doses to be studied is bounded between placebo (the zero dose), and MTD. Although the true underlying efficacy dose-response relationship may not be monotonic, it may still be reasonable to assume monotonicity of efficacy dose-response within the range between placebo and MTD. At least it may be assumed that the dose at or close to MTD could still deliver an efficacy response that is significantly different from placebo.

Figure 4.1 distinguishes between individual dose-response relationships—the three thinner curves representing three different individual subjects—and the single, thicker population average dose-response relationship. Because of inter-subject variability, different subjects may respond to the same drug differently (Holford and Sheiner 1981).

In Phase II clinical development of a medicinal compound, the dose-response relationship is typically viewed as a population dose-response (instead of individual dose-response) and such a relationship could be estimated based on responses observed across the entire patient population being studied. In addition, a dose-response relationship can also be viewed on an individual basis. When this is the case, the response to a given dose of the study product from a particular individual can also be contemplated. The efficacy response of a particular patient to the particular dose of the test product could be a continuous function or, alternatively, it could be a step function.

In the case of a continuous relationship within an individual patient, it can be thought that, as doses of the candidate product increase, the amount of protein within the human body responding to the product increases. For example, in the development of a beta blocker for reducing systolic blood pressure, more drugs may

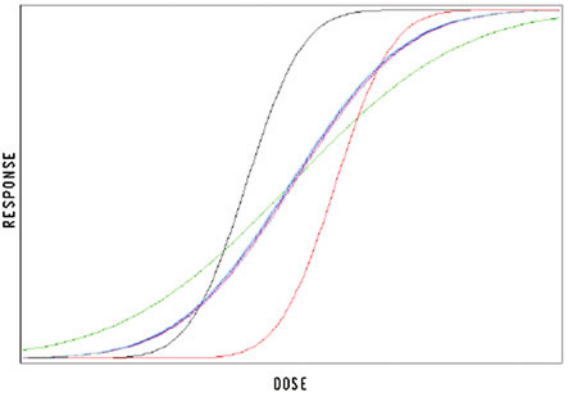

Fig. 4.1 Individual dose response and population dose response

be able to inhibit more beta agonists. Accordingly, the symptom reduces in an incremental fashion. Under this assumption, for each individual patient, the dose and the response follow a monotonic relationship.

Alternatively, the individual patient's response to increase of doses could be a step function, where an underlying threshold exists in the human body. When the amount of doses in the body is below the threshold, the patient does not experience any response. Once the dose increases to be over the threshold, then the patient responds.

In most cases when the monotonicity assumption is considered, the implied dose-response relationship tends to be a population dose-response. If people are skeptical about a monotonic population dose-response assumption, can the assumption of monotonic individual dose-response be acceptable? If so, then the non-monotonic population efficacy dose-response may be explained using the individual monotonic dose-response. For example, if the within patient dose-response follows a step function, and the threshold of each patient is different, then the observed population dose-response may reflect a random proportion of responders at various doses. Moreover, if the test product is indeed efficacious, and if the sample size of each treatment group is large enough, then some of the doses could demonstrate a statistical difference from placebo, although no apparent dose-response monotonicity can be observed. In cases where responses do not increase as doses increase, this can still be considered as monotonic dose-response relationship.

Given the previous discussion, we believe it may still be useful to make the monotonicity assumption at the population level in designing the PoC trial. Under this situation, the choice of MTD as the dose of test product in comparison with a placebo control could still be considered a reasonable choice. Again, even if the true (underlying) dose-response relationship may not monotonic, it is likely that the MTD may be able to deliver sufficient efficacy response to support PoC.

4.6 Agreement on a Delta

One of the most difficult questions in designing any clinical trial is the agreement of a treatment difference (δ) between the test product and placebo. Sometimes the discussion about δ may take a long time and may involve in team members from many scientific background. It is well known that a treatment difference between the test product and the placebo control needs to be "clinically meaningful." However, during the discussions about proposing a δ to be used for study design, a minimally clinically important difference (MCID) is very difficult to achieve. The major challenge is that it is really hard to know how much improvement in the primary efficacy measure delivered by a useful treatment can realistically benefit the patients with the condition or the disease under study.

For example, in the development of a new test product to treat pain, and the primary endpoint is a 11-point self-report pain scale, from 0 being no pain to 10 being the most severe pain. In a two-week clinical trial, patients are randomized to test product and placebo, and the mean reduction in pain from baseline to week 2 is compared between the two treatment groups. The question is which δ should be used for designing such a study.

Suppose the mean reduction on 0–10 pain scale is calculated as the week 2 measurement minus the corresponding baseline measurement. Then both treatment groups would observe negative mean changes if there is benefit for both groups. Let the treatment difference be obtained from the test product mean subtract the placebo mean, the delta would be expected to be a negative value. The question becomes what values of δ are considered clinically meaningful. From a sample size calculation point of view, these various δ values could translate into large differences in sample sizes.

Several instruments have their clinically important differences or minimum clinically important differences (MCID) (Katz et al. 2015; Rosen et al. 2011; Osoba et al. 1998). One example is the reduction in blood pressure in an anti-hypertensive trial—it is generally accepted that a mean reduction of 5 mmHg in comparison with placebo would be clinically meaningful. Another example is the erectile function domain of the International Index of Erectile Function for men with erectile dysfunction, with a minimum clinical important difference of 4 points (Rosen et al. 2011). A third example are the domains on the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire, with such a minimum given as 5 points. A fourth example pertains to the treatment of rheumatoid arthritis (RA). The widely accepted primary measure is ACR20, which is a 20% reduction in symptoms based on a composite of seven symptom measures proposed by the American College of Rheumatology (ACR). It is generally expected that in a short-term RA trial, there would be about 30% placebo responders. Hence a clinically meaningful difference of 15–20% treatment improvement would imply an observed ACR20 response of 45–50% can be expected from the test treatment, if the test product is truly efficacious in treating patients with RA. A fifth example is the 0–10 numeric rating pain scale with its two-point reduction (or more extreme) considered as clinically important (Farrar et al. 2001).

However, for most of disease conditions, an MCID is very difficult to agree upon (King 2011; Copay et al. 2007; Hays et al. 2000). Generally speaking, there are three approaches of proposing a treatment difference for sizing a study. The most frequently used method is to learn from other products. One example is to look at the treatment differences observed from the approved products developed by other sponsors. In the US, this could be obtained from the Summary Basis of Approval (SBA) issued by FDA. In the European Union, the information could be found in the European Public Assessment Report. One other way of learning about postulating a reasonable treatment difference is by studying compounds developed within the sponsor themselves. Such a learning could be obtained from previous experiences of a similar product from the same class of the new candidate product or even a product from a different class.

The second approach of postulating a reasonable treatment difference is based on prior information of the product being developed—the data collected from pre-clinical studies or from Phase I studies. Sometimes biomarker information can be used to help with this effort. For example, in the development of a beta blocker, the amount of inhibition for beta agonists in Phase I studies can be useful. In certain situations a combination of both approaches (using information from other products and prior information of the same product) are considered in reaching an agreement of a delta (treatment difference) to be used for sample size calculation.

The third way of guessing a delta is interesting, which also happens in practice, not infrequently—to start with a reasonable sample size and then reverse the sample size calculation to identify a feasible delta. This happens when there is an entirely new indication or a new endpoint, where there is no previous experience. This situation may also happen when patient recruitment is slow or when the budget is limited. When this is the case, team members often spend considerable time to discuss pros and cons about various choices of δ. A final way to reach agreement is based on a compromise on two or three of these approaches.

One question associated with choice of delta concerns variability corresponding to the primary endpoint. In most disease areas, the standard deviation is known and the variability tends to be stable. For example, in a four-week treatment of patients with hypertension, the standard deviation of changes in blood pressure within a treatment group is about 10–15 mmHg. In the case of a binary endpoint, based on the binomial distribution, if the proportion of responders corresponding each treatment is known, then the variability can be calculated with the given sample size. If we know the proportion of patients who respond in each of the two groups, as well as the alpha and power, we can get the sample size needed. Once the discussion of delta is completed, a reasonable standard deviation would be used to simplify these two quantities and obtain the effect size. This is easier to understand when the outcome is continuous as the effect size is the delta divided by standard deviation (signal-to-noise ratio).

Agreeing on a clinically meaningful treatment difference from placebo control is a very difficult challenge in the early Phase II clinical trial design. Such is one of the greatest difficulties at this stage. However, without a delta, it would almost be impossible to calculate the sample size for a clinical trial. Once the delta is obtained,

the next natural step is to discuss with the team regarding types of risks the team is willing to take, namely, selection of a Type I error rate and a Type II error rate.

4.7 Choice of Alpha and Beta

In clinical development of new products, it is well known that alpha is the probability of making a Type I error (i.e., the probability of essentially developing a placebo). Beta is the probability of making a Type II error, or the probability of stopping development of a potentially good product. In most of the Phase III development programs, alpha is set at one-sided at 0.025, because this is required by regulatory agencies; this quantity is not negotiable. Alpha can be thought of as a regulatory risk. If a product without efficacy is approved for the general patient population use, then millions, or tens of millions of patients could use a product which does not deliver efficacy, but could cause adverse events or other safety concerns.

Beta can be thought of as the sponsor's risk. If a potentially good product was stopped for further development because of a wrong decision, then the sponsor's investments on this candidate would have been wasted, resulting in opportunity cost. For this reason, regulatory agencies do not require any particular beta to be used in any phase of the clinical studies. Hence the choice of beta is usually left for the sponsor to determine. Usually in Phase III, sponsors tend to choose a smaller beta (or a higher power) for study designs so that a larger sample size will increase the likelihood of providing a smaller *p*-value. A typical Phase III study tends to be with a design of beta being 20% or less.

However, choice of alpha or beta in Phase II trial designs can be flexible. The two main reasons are that (1) decision after a Phase II clinical trial is a sponsor decision, not a regulatory decision and hence the sponsor gets to select the level of risks they are willing to take; and (2) there is insufficient amount of evidence at Phase II for the sponsor to commit a huge amount of investment. Accordingly, under a reasonable or affordable sample size, the alpha or beta may have to be larger.

When designing a Phase II study, if the δ has been determined, and the sample size is limited, then the only choice left for the project team would be whether to take a larger risk of making a Type I error or a larger risk of making a Type II error, or both. This choice is critical, because a discussion about probability of making a wrong decision should take place at the design stage, not after the study is unblinded and results are ready. From a frequentist point of view, the probability of observing a head or a tail of a coin toss can only be contemplated before the coin is landed. After the coin is landed, the observation is either a head or a tail. There would be no uncertainly involved as the outcome is known and certain. In clinical trials, the coin is landed when the blind is broken. At that time, clinical data are ready for statistical analysis, and a *p*-value can be calculated. It is important to understand that such a *p*-value is an observed *p*, and there is no uncertainty associated with it.

The underlying truth—whether the test product is in fact efficacious or not—is still unknown. Yet a Go or NoGo decision will have to be made. What can be concluded at this stage is that if the observed *p*-value is small (less than alpha), and a Go decision is made, then the probability of making a Type I error under such a decision is controlled at a level alpha. If an observed *p*-value is large (*p* greater than alpha), and a NoGo decision is made, then the probability of making a Type II error is maintained at level beta.

In Phase III, the selection of beta reflects the level of commitment of moving the candidate product forward (a smaller beta implies a stronger commitment). It also represents the degree of investment the sponsor is willing to spend. On the other hand, in Phase II, the choice of beta may need to be balanced with the choice of alpha. In most of situations, beta is set at a higher level than alpha. For example, a typical alpha is 0.05 (either one-sided or two-sided), while a typical beta is 0.2 (indicating 80% power). These settings do not necessarily have to be the case in all Phase II designs. In practice, Phase II studies are designed with a one-sided alpha of 0.0125, 0.025 and 0.05 or sometimes at one-sided alpha of 0.075 or 0.1. A beta value of greater than 0.2 is rare. But beta values of 0.15, 0.1 or 0.05 are not uncommon.

It frequently takes the project team to discuss and agree with the level of risk the team is willing to take—either to give up a good product or to develop a placebo. From a frequentist's point of view, it is very important to note that these discussions should take place at the study design stage, because after the results are known, the coin has landed, and it would be too late to discuss probabilities or risks.

4.8 Sample Size Considerations

In the design of a clinical trial (PoC, dose-ranging, Phase III, or any other clinical efficacy trials), a typical first step is to clearly articulate a clinically important question. Next, team members collaborate to identify the primary endpoint, to agree on the treatment difference (δ), and to choose Type I and Type II error rates. A statistical hypothesis is then formulated and, finally, the sample size estimate is calculated based on this information.

Whenever sample size calculation is needed for a given study, the first point is to understand the hypothesis to be tested. In the process of designing a Phase II PoC clinical trial, the statistical hypothesis associated with such a trial is typically a simple hypothesis with a one-sided test. In general, there are only two treatment groups, with a placebo group compared against the MTD of the test product. The hypothesis can be expressed in a straight-forward fashion as responses from the test treatment against responses from placebo. For a comparison based on treatment means, where higher sores are more favorable, then the hypothesis can be written as

$$H_0 : \mu_T = \mu_P \quad \text{versus} \quad H_0 : \mu_T > \mu_P \tag{4.1}$$

where μ_T is the test treatment group mean and μ_P is the placebo treatment group mean. In other words, if mean responses of the test treatment group is statistically significantly greater than the mean responses of the placebo group, then the concept is considered proven.

Once the statistical hypothesis, the clinical difference, and the corresponding alpha and beta have been determined, sample size for a given clinical trial can be calculated. Here a general formula for sample size calculation is provided, for instance, it is used to help calculate sample sizes under a dose-ranging study. For PoC, the number of treatment groups reduces to two, and the number denoted by k in the following formula reduces to be one ($k = 1$, for a PoC trial). For PoC with more than 2 groups, see Chap. 6. This formula will be re-visited in the next chapter (Chap. 5) where dose-ranging trials will be discussed. Based on the above discussion, the sample size of a Phase II (either PoC, or dose-ranging) study can be calculated using the following formula (Chang and Chow 2006):

$$n = \left[\frac{(z_{1-\alpha} - z_{1-\beta})\sigma}{\delta}\right]^2 \sum_{i=0}^{k} \frac{c_i^2}{f_i}$$

where

n	is the total sample size of all treatment groups combined,
$z_{1-\alpha}$ and $z_{1-\beta}$	are the critical values of the standard normal distribution corresponding to Type I and Type II error, respectively,
δ	is the treatment difference between test product and placebo,
σ	is the standard deviation,
f_i	is the fraction of number of patients in treatment group i,
c_i	is the contrast coefficient that is associated with treatment group *i.*

In the case of two group PoC, then the coefficients are −1 and 1.

It is assumed to have k dose groups, plus placebo, where placebo is denoted as the 0 dose group. Again, in the case of designing a two-group PoC clinical trial, $k = 1$.

References

Chang, M., & Chow, S. C. (2006). Power and sample size for dose response studies. In N. Ting (Ed.), *Dose finding in drug development* (pp. 220–241). New York: Springer.

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal, 7,* 541–546.

Farrar, J. T., Young, J. P., LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain, 94,* 149–158.

Hays, R. D., & Woolley, J. M. (2000). The concept of clinically meaningful difference in health-related quality-of-life research, how meaningful is it? *Pharmacoeconomics, 18*(5), 419–423.

Holford, N., & Sheiner, L. (1981). Understanding the dose-effect relationship: Clinical application of pharmacokinetic-pharmacodynamic models. *Clinical Pharmacokinetics, 6,* 429–453.

Ivanova, A. (2006). Dose finding in oncology—Nonparametric methods. In N. Ting (Ed.), *Dose finding in drug development* (pp. 49–58). New York: Springer.

Katz, N. P., Paillard, F. C., & Ekman, E. (2015). Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions. *Journal of Orthopaedic Surgery and Research, 10,* 24.

King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Reviews of Pharmacoeconomics & Outcomes Research, 11*(2), 171–184.

Osoba, D., Rodrigues, G., Myles, J., Zee, B., & Poter, J. (1998). Interpreting the significance of changes in health-related quality-of-life scores. *Journal of Clinical Oncology, 16*(1), 139–144.

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

Rosen, R. C., Allen, K. R., Ni, X., & Aruujo, A. B. (2011). Minimal clinically important differences in the erectile function domain of the international index of erectile function scale. *European Urology, 60*, 1010–1016.

Chapter 5
Design of Dose-Ranging Trials

5.1 Background

Typically, Phase I clinical trials are designed to evaluate escalating doses in order to find the maximum tolerable dose (MTD) so as to establish the upper anchor of doses for future development of new medicinal products. After Phase I, one of the most important objectives for Phase II clinical development is to answer the Go/NoGo question—proof of concept (PoC). For new drugs treating non-life-threatening diseases, the next important objective is to establish the "dose range." A range of doses can be identified by the high dose and the low dose, and the high dose is limited by MTD. As such, a key Phase II objective is to find a low dose of this dose range. Usually this low dose is considered as minimum effective dose (MinED). Simply put, Phase I is moving doses upward to find MTD; Phase II is moving doses downward to find MinED.

After the concept is proven, the most important next step is to explore and recommend the commercial doses to be tested in large Phase III confirmatory clinical trials. PoC is usually faster and cheaper as it only requires a well-tolerated dose of test therapy plus the placebo (control) group. A dose-ranging study, however, needs multiple doses of the test therapy to characterize the dose-response relationship. Therefore, a classical Phase II development program usually consists of small-scale PoC trials followed by moderately sized dose-ranging studies.

Conceptually, the efficacy of a candidate product increases as the dose increases (Fig. 5.1). The left curve plotted in Fig. 5.1 can be viewed as an efficacy–response curve, with the response (y-axis) representing some measure of product efficacy. In addition to efficacy response, there are also toxicity dose–response curves. We will assume, as shown in Fig. 5.1, when dose increases, both efficacy and toxicity increase. Based on these curves, the maximum effective dose (MaxED) and the MTD can be found: MaxED is the dose above which there is no clinically significant increase in pharmacological effect or efficacy; MTD is the maximal dose acceptably tolerated by a particular patient population.

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_5

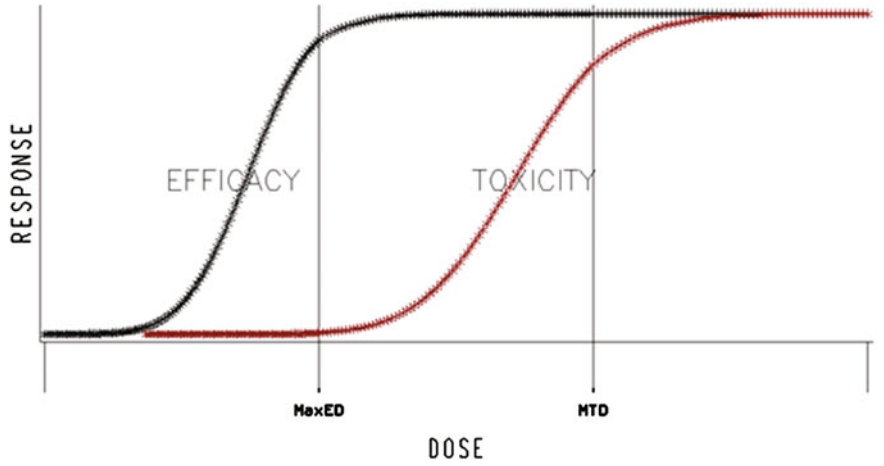

Fig. 5.1 Efficacy and toxicity dose response curves

In certain products, the efficacy curve and the toxicity curve are widely separated. When this is the case, there is a wide range of doses for patients to take; that is, as long as a patient receives a dose that is greater than the minimum effective dose, and below a toxic level, the patient can benefit from the efficacy while avoiding certain toxicities from the test product. However, for other products, the two curves may be very close to each other. Under this situation, physicians have to dose patients very carefully so that while benefitting the efficacy, patients do not have to be exposed to potential toxicity from the candidate product.

The area between the efficacy and the toxicity curve is known as the "therapeutic window." One way to measure the therapeutic window is to use the "therapeutic index (TI)." TI is thought of as the ratio of TD_{50}/ED_{50}, where TD_{50} is the dose that produces toxicity in 50% of population and ED_{50} is the efficacious dose for 50% of the population. Clearly, a product with a wide therapeutic window (or a high TI) tends to be preferred by both physicians and patients. We want the drug to have sufficient efficacy with the smallest dose in order to lessen or minimize side effects (even when the dose is on the low side). If a test product has a narrow therapeutic window, then it will need to be developed carefully. Caution is needed should the medicine be approved and prescribed to the target population.

Frequently, products developed for life-threatening diseases (e.g., cancer) tend to have narrow therapeutic windows. In contrast, products developed to treat chronic diseases (e.g., high blood pressure, depression, high cholesterol, or arthritis) are required to have wide therapeutic windows, because these products are frequently used in older patients, or patients with other concurrent diseases, and sometimes used in children.

In clinical development, Phase III trials are designed to confirm product efficacy. Results obtained from Phase III studies provide evidence to help regulatory agencies make decisions to approve or not to approve the candidate product.

Therefore, it is critical that dose-ranging trials from Phase II deliver the information to help with Phase III design. In other words, if the appropriate dose or doses for product registration is not included in Phase III trials, it is likely that the compound will miss the opportunity of being approved for marketing. Hence information obtained from Phase II dose-ranging trials need to provide sufficient support for dose selection in Phase III study designs.

5.2 Finding Minimum Effective Dose (MinED)

Minimum effective dose (MinED) is a crucial concept in product development. Unfortunately, there is no universally accepted definition of this term (Thomas and Ting 2009). Finding MinED is critical because it is generally believed that toxicity increases as dose increases. Hence a lower dose that is effective implies it could be safer than higher doses.

Clinical results obtained from Phase III studies are used to support product registration. Based on these data, if a dose is efficacious and safe, then that dose can be approved, and the product label specifies the dose or the range of doses the product is approved for. Nevertheless, after approval, if excessive adverse events are observed from the low dose of the product, then a question as to whether this low dose is low enough could be raised. At this time, it would be very challenging to ask what is the MinED for this product. Therefore, it is critical to find MinED in Phase II, before it is too late.

Drug price in US is not directly associated with dosage. However, in many other countries, the price of a medicinal product is linked to its dose. In other words, if the dose of the product is higher, then it is more expensive. Now suppose a certain product is approved in such a country and the price is set based on the lowest approved dose. If later it was detected that the MinED for the product is, in fact, lower than the approved dose, then a dose reduction is needed at a post-market stage. Accordingly, the price of this product at a lower dose may have to be further lowered as well. When this happens, there is a huge impact on the earnings from such a product. Therefore, from a marketing strategy point of view, it would be very critical to identify the MinED before the product enters its Phase III development.

MinED Finding is performed in the practice of Phase II trial. As discussed previously, if MinED was not identified at Phase II, it would be very difficult and very expensive to find the MinED at a later stage of product development. In fact, Phase II studies may be the only opportunity to characterize the relationship between doses and product efficacy. Therefore, it is vital to pay more attention to Phase II and to carefully design dose-ranging clinical trials during this critical phase of clinical development for a new product.

Although there is a definition of MinED from ICH E4—"MinED is the **smallest dose with a discernible useful effect**," the interpretation and implementation of such a definition varies a lot in practice. The first question is the understanding of "a useful effect." It is difficult to achieve consensus about the smallest magnitude of an

effect that is useful. One approach is to follow a well-established treatment effect such as lowering 5 mmHg of blood pressure in developing an antihypertensive product. Another is to consider the concept of minimal clinically important difference (Farrar et al. 2001; Katz et al. 2015; Osoba et al. 1998; Rosen et al. 2011) the magnitude of the treatment effect above and beyond placebo effect is large enough to be perceived by a subject, described in the previous chapter.

The implementation of a "discernible" effect could be even more difficult. Typically a "discernible" effect could be referred as a "statistically significant" difference from control. Under this interpretation, various authors recommend different ways of calling a dose as MinED based on the difference in treatment effect from a given dose as compared with placebo. Some researchers suggest the use of a dose-response model to help estimate MinED (Fillon 1995; Pinheiro et al. 2006). In the modeling approach, a treatment difference (δ) against placebo is proposed and from the model, if there is a dose achieves such a δ, then this dose is considered as the MinED. Others (Hochberg and Tamhane 1987; Wang and Ting 2012) suggest the use of pairwise comparisons. Under pairwise comparisons of each dose versus placebo, a point estimate and an interval estimate can be calculated. Depending on the definition of MinED, these estimates are used to help proposing a dose as MinED. Even among those authors recommending use of pairwise comparisons to find MinED, there could be many differences, too. For example, some may argue that multiple comparison adjustment could be necessary, others may not. Whether or not multiple comparison is needed, the understanding of "a dose that delivers treatment effect which is different from placebo" could still vary substantially, depending on individual's point of view.

For example, let the useful effect be denoted as δ; then, for a given dose, a point estimate and a confidence interval can be constructed for δ. If a positive value indicates a benefit effect, then the treatment response of the given dose of test product subtracted from the placebo response is expected to be positive. Hence it is hoped that the point estimate of the treatment difference could be as much as δ. This leaves a wide range of interpretations—for MinED, should that dose be such that the point estimate is greater than δ? Should that dose be such that the lower confidence limit be greater than 0? Should that dose be such that the upper confidence limit be greater than δ? Or should that dose be such that the lower confidence limit be greater than δ (McLeod et al. 2016)?

The implementation of "the **smallest dose with a discernible useful effect**" is fraught with various interpretations and complexity, making the definition difficult to implement unambiguously. In practice, one feasible way of finding a MinED in a dose-ranging clinical trial is to find a dose that is lower than MinED. Regardless of which practical definition the project team uses to identify MinED, suppose there is a dose deemed to be "not efficacious," and another dose that is higher could be considered to be "efficacious." Then it can be considered that the non-efficacious dose is below MinED and the efficacious dose is above MinED. From a practical point of view, this information would usually be sufficient to help the team to design Phase III clinical trials.

Of course in the selection of Phase III doses, many other aspects will also have to be considered: such as safety or toxicity findings, pharmacokinetic properties, formulations, and so on. But the information about MinED would be one of the most critical deliverables from Phase II in supporting any Phase III development plans.

5.3 A Motivating Example

In a clinical development program for a test drug to treat osteoarthritis, the first dose response study included placebo, 80, 120, and 160 mg of the test drug (Ting 2008). Results from this study indicate that all three doses are efficacious (Fig. 5.2a). These results also show that 80 mg may have already been at the high end of the efficacy dose-response curve. A second study was designed to explore the 40 mg dose, and results from this second study are given in the middle panel of Fig. 5.2b. With these findings, the project team started Phase III studies using a dose range of 40–120 mg to confirm the long-term efficacy of this test drug.

An End-of-Phase II meeting with the Food and Drug Administration (FDA) was held after the long-term Phase III study started. FDA commented that the minimum effective dose (MinED) was not yet found. Hence the project team designed a third dose-response study to explore the lower dose range. Doses included in the third study are 2.5, 10, and 40 mg. As shown in the bottom panel of Fig. 5.2c, the third study successfully established the dose-response relationship. Based on these results, it is clear that the Phase III studies were designed with a range of doses that are too high. The impact of this process can cause a major delay in the development program and considerable waste of resources.

5.4 How Wide a Range of Doses to Study?

The key lesson learned from this osteoarthritis drug development program is that a very low dose was not explored in early Phase II. In other words, if a dose at or below 10 mg was explored in the first study, a narrower dose range could have been established using a second study. Then the Phase III studies could have been designed with the appropriate dose range.

In general, the most difficult challenge in designing the first dose ranging trial in developing a new intervention is the selection of doses to be included in this design. Just like designing a PoC trial, the dose-ranging study is designed with two very important assumptions:

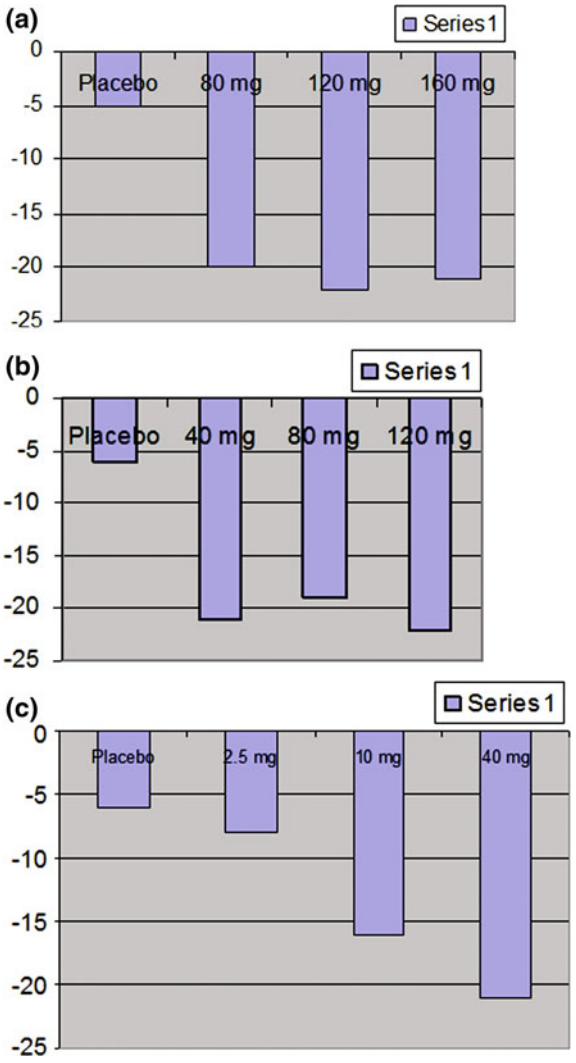

Fig. 5.2 Dose response results from the three studies of osteoarthritis drug

(1) The maximal tolerated dose (MTD) was correctly estimated from earlier trials and
(2) Efficacy response is monotonic to ascending doses.

However, it is unclear what range of doses to choose for efficacy activities. At this stage, the best guidance available is data obtained from pre-clinical experiments or from Phase I biomarkers. Even under the assumptions that the MTD was correctly estimated and there is a monotonic efficacy dose-response relationship, there can still be many possible shapes of dose-response curves.

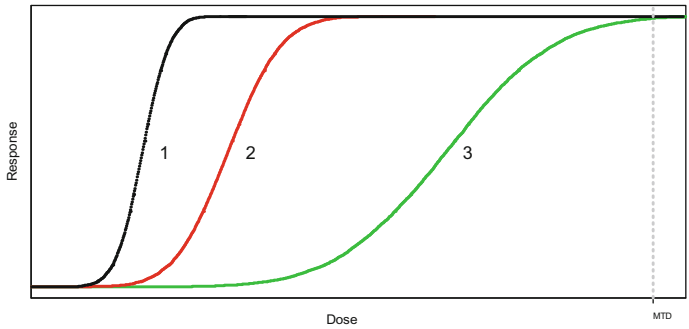

Fig. 5.3 Several possible dose response curves

Figure 5.3 presents an example of three dose-response curves. The three curves on this figure represent dose projections based on three animal models from non-clinical studies. One curve represents a dog model, another is a mouse model, and the third is from a rabbit model. The vertical dashed grey line (at the right side of Fig. 5.3) represents the maximum tolerated dose (MTD). The question then becomes, what range of doses should be considered in the upcoming dose-ranging clinical trial? From this perspective, the primary challenge in the dose-ranging trial design is to consider which range of doses needs to be studied.

5.4.1 Definition of Dose Range in a Given Study

Dose range in a clinical trial design is defined as the ratio calculated as the highest dose divided by the lowest dose included in the same design. For example, if the doses in trial A are placebo, 20, 40, and 80 mg, then the dose range is 4 (= 80/20). If the doses in trial B are placebo, 0.1, 1 and 10 mg, then the dose range is 100 (= 10/0.1). Although the doses used in trial A are higher than doses used in trial B, trial B has in fact a much wider dose range—25 times wider than trial A.

Given the discussion in the previous sections about MinED, it is critical to realize that in the first dose-ranging design, a wide dose range need to be studied. The high dose is typically set at MTD or a dose below MTD to avoid any potential toxicity. The real design challenge is to choose the lowest dose for this dose-ranging trial. If there is an interest to identify a dose that is below MinED, then the difficult question would be "how low would it be low enough?" Once the low dose is selected for the design, a dose range can be calculated as the ratio of the highest dose divided by the lowest dose.

5.4.2 Binary Dose Spacing

Hamlett et al. (2002) propose to use a binary dose spacing (BDS) design for dose allocation. If the trial includes two dosing groups and a placebo, then the BDS design identifies a mid-point between the placebo and the MTD and allocates one dose above and another below the midpoint. If the trial includes three dosing groups and a placebo, then the BDS design identifies the high dose as described in the two dose case. A second midpoint is then identified between the placebo and the first midpoint, and a low dose is selected below the second midpoint and the medium dose is selected between the two mid-points. The BDS design iteratively follows this strategy with more dosing groups. The BDS design employs a wide dose range, helps identify the MinED, employs a log-scale dose spacing strategy avoiding allocating doses near the MTD (i.e., identifies more doses near the lower end), and is flexible and easy to implement.

5.5 Frequency of Dosing

In designing dose-ranging clinical studies, we need to know how often should a patient take the test drug (e.g., once a day, twice a day, or dose every 4 h during the day). This is a question about dosing frequency, and it is usually guided by the Phase I pharmacokinetics/pharmacodynamics (PK/PD) findings. One of the PK parameters is the half-life of a drug (this is a PK parameter that estimates how long will it take to excrete 50% of test drug out of the body). The estimated half-life helps to explore how long the drug will stay in human body. Using this information, we can propose a dosing frequency to be used in a dose-ranging study design. In certain cases, we may study more than one dosing frequency in a single study. When this is the case, a factorial design (dose, frequency, dose*frequency) could be considered.

However, in some drugs, the PD response may be different from the PK response. Note that PD measures how the drug works in human body while PK measures how the body acts to the drug. In the example of developing a beta blocker, the beta receptors could be served as one of the PD markers. When this is the case, based on the PK half-life data, we may believe that insufficient drug in the body even after several hours of dosing and there may still be enough drug in the tissue to help with PD responses. On the other hand, the PK may indicate that there are still plenty of drug in the body, but these drugs may not cause any effective PD responses. In some drugs, the concentrations for PD activities can be very different from that for PK activities. Hence the dose frequency purely derived from PK may either overestimate or underestimate the concentration needed for PD responses. In some cases the best dose frequency may be derived in later phases of the drug development.

Another important guiding principle in selecting dosing frequency is based on the market assessment. For example, if the market requires a once-daily dosing treatment, but the drug candidate under development has a twice-a-day PK profile, then some formulation change may be necessary. Figure 5.4 presents time-concentration curves for this situation. The horizontal line that is directly above the x-axis represents the efficacy concentration level (often based on PD information). Theoretically speaking, we need to keep the drug concentration staying above this line all the time for the drug to work. In order to achieve this concentration, two strategies are possible: (1) dose the subject twice a day (BID) with low dose (Fig. 5.4); or (2) double the BID dose, and treat the patient with this high dose once a day (QD) (Fig. 5.4). Note that in the first few hours post dosing, the high dose may result in a very high concentration which could potentially cause severe adverse events. When this is the case, re-formulation of the drug may be needed so that, for dosing once a day, the Cmax (which is a PK parameter that estimates the maximal amount of drug observed in plasma at any given time) would not be too high, while the efficacy concentration can be maintained throughout a 24-h period.

In designing the first few Phase I trials to study PK, there is very limited information about how the human body will metabolize the drug candidate. At this stage, data observed from pre-clinical studies and animal experiments on this candidate are used to help guide designing these Phase I trials. In case the drug candidate belongs to a certain drug class where other drugs of the same class were already on the market, information obtained from these other drugs can be used to help guiding the study designs for the drug candidate being studied.

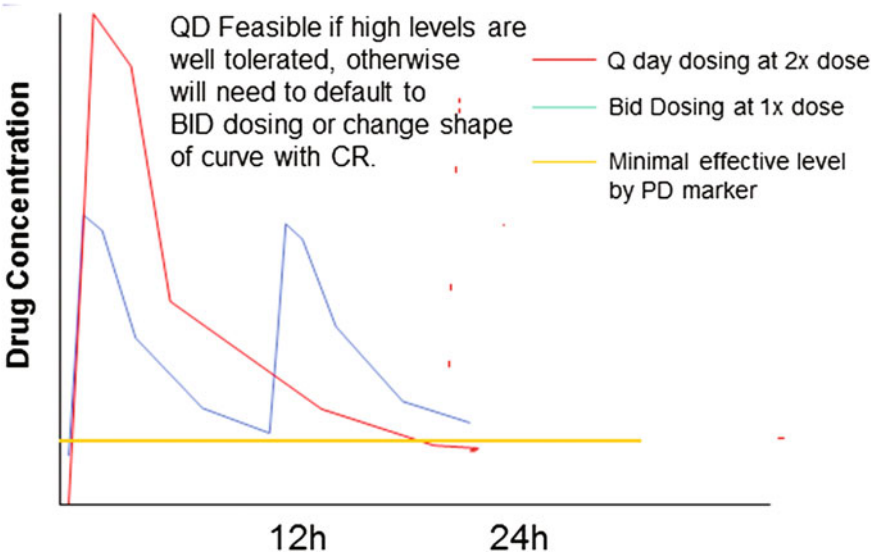

Fig. 5.4 Once a day versus twice a day dosing (QD is once a day, BID is twice a day)

During the development of a compound, sometimes reformulation may be needed. There can be many different reasons why there is a need for drug reformulation, including to help absorption and to change the half-life. It is critical to understand that after reformulation, the PK/PD properties of the product candidate are different from what they were prior to reformulation. Hence all of the dosing information and drug regimen obtained from studies before the reformulation will need to be changed and re-studied. This change can potentially cause major re-work. Re-work in drug research and development delays the development process and results in additional amount of investment.

In studying the PK/PD relationship, we should realize that the main point is whether Cmin (Cmin is the PK parameter that estimates the minimal amount of drug observed in plasma), Cmax (the maximal concentration), or AUC (area under the curve) drives the PD. This area is fertile for collaboration between statisticians and pharmacokineticists. Models based on prior trial data (e.g., from pre-clinical data, clinical data of the drug candidate under study, other compounds of the same class) can be developed to inform the decision.

5.6 Parallel Controlled Fixed Dose Designs

Typical dose-ranging trials are designed with parallel dose groups plus a placebo control. Patients are randomized to each treatment group and receive study medications for the entire duration of the study. For example, in a four treatment group design with four weeks duration, the doses are placebo, low dose, medium dose and high dose, then a patient is randomized into one of these four groups and will receive the same dosage of study medication from after randomization to week 4. A parallel design indicates that at the time a patient is recruited, the opportunity of randomizing into each dosing group is dictated by the study design, and that there is no requirement for any study dose to be before, or after another study dose.

Dose-escalation designs are frequently used in Phase I. In a dose-escalation design, a group of subjects started with the lowest designed dose. After these subjects complete the lowest dose and there is no safety concern, then a different group of subjects will receive a higher dose. If this dose is well tolerated, then a third group of subjects will receive an even higher dose. This process continues until the stopping rule is met. A typical stopping rule would be some kind of pre-specified adverse events. In oncology, the stopping rule could include a dose limiting toxicity (DLT). In other words, if several subjects experiencing DLT's, then it can be claimed that the MTD was exceeded. Dose-escalation designs are very useful in finding the MTD. However, since the duration of each treatment period is relatively short, these designs are not used in clinical efficacy trials. Phase II dose ranging trials are designed to study drug efficacy and, in most of therapeutic areas, drug efficacy can only be observed after a reasonable treatment period. Hence dose-escalation designs are not used in dose-ranging trials.

A fixed-dose design and a dose-titration design are in contrast to each other. In a fixed-dose design, once a patient is randomized to a dose group, the patient would take the same dose of study product throughout the complete dosing period. In a dose-titration design, in contrast, a patient is randomized to a dose regimen with a starting dose, where the dose for a patient can be changed over time. In a titration design, patients are randomized to regimens, instead of doses. The regimen could be a 10–20 mg test drug versus placebo. In this case, a patient could be randomized to test drug or placebo. If the patient is randomized to test drug, he or she would be dosed at 10 or 20 mg of test drug, depending on study design. If a patient is randomized to placebo, then this patient may receive 10 mg matching placebo or 20 mg matching placebo, according to the double-blind study design.

Dose-titration designs include forced titration, and response guided titration. For example, in an eight-week study with forced titration of test drug against placebo, with 10 and 20 mg dosages, then the design may dictate that every patient starts with 10 mg of test drug or matching placebo. Then after two weeks of 10 mg (or placebo) double-blind treatment, every patient will be titrated up to 20 mg, until end of treatment (week eight). On the other hand, a response-guided titration design (again, two treatment groups with test drug against placebo, and eight weeks duration) would randomize patients into one of the two treatment groups, and every patient starts with 10 mg (or matching placebo). Then, depending on patient response, if the efficacy is not sufficient, then the treating physician may consider titrate the dose up to 20 mg. However, after the patient on the 20 mg dose, if there is safety concern at this dose, then the physician may down titrate the dose back to 10 mg. Hence dose-titration designs can be viewed as two different types—forced titration or response-guided titration.

There are some advantages of a titration design. For example, a study with this design will allow a patient to be treated at the best dose for that specific patient; this dose allocation feature reflects the actual medical practice. However, the disadvantage of a titration design is the difficulty in data analysis. For example, if a patient responded to the test drug after doses are titrated up, it is unclear whether the higher dose delivered the efficacy or the accumulation of lower doses caused the response. In titration designs with multiple treatment groups, there may be overlapping doses—for example, one treatment group is 10 mg escalating to 20 mg, while another group is 20 mg escalating to 40 mg. When this is the case, it is difficult to make inferences about the 20 mg dose group.

In some rare cases, instead of a dose-response study, a concentration response study could be considered. A concentration response study assesses efficacy and safety measurements observed from subjects according to the plasma concentration of the study drug but not the doses of the study drug. There are many practical limitations in using this type of designs. These limitations include, among others, how to blind the patient and the physician and also how and when to measure the blood concentration. Because of the issues with dose-titration designs and concentration response designs, the parallel, fixed-dose designs are, in general, the more commonly used designs for dose-ranging studies. Therefore, in many of the

dose-response studies, patients are randomized to a few fixed-dose groups and are compared with one or more control treatment groups.

In Phase II dose ranging clinical trials, the primary objective is to evaluate drug efficacy of each studied dose, and hence the most useful design is the parallel, controlled, fixed-dose design. In these designs, each patient is randomized to a pre-specified dose and followed for the designed study duration. During the entire study period, patients received the same dosage of the double-blind medication. Results obtained from these studies can be used to help understanding drug efficacy associated with each studied dose. It is important to note that dose-escalation design, dose-titration design or concentration response design cannot be used for efficacy dose-response purposes.

5.7 Number of Doses and Control Groups

As discussed in the previous sections, a wide dose range is critical in the Phase II dose-ranging studies (Yuan and Ting 2014). Once MTD is estimated from earlier studies and a wide dose range is considered, the natural questions to think about are how many doses should be included in the study and whether a placebo-control group is sufficient or an active control group is needed.

A single high-dose plus a placebo-arm design may be able to demonstrate PoC but cannot adequately characterize the dose-response relationship. Any attempt to interpolate a dose-response relationship between the placebo and the test dose would require very strong assumptions, which often times are not realistic. Therefore, a dose-ranging study typically needs several test dose levels. Generally speaking, it would be desirable to have more than two test doses plus one placebo arm in defining the dose-response relationship and estimating the difference of the test doses versus placebo. One very commonly used design is a four-group study with placebo, low-dose, medium-dose and high-dose groups of the drug candidate under development.

Some authors suggest that more doses could help (Krams et al. 2003)—that is, in the first dose-ranging study, adding more doses to the study design. However, the number of doses to be used in a dose-ranging trial is usually limited by practical and logistical considerations, such as available formulations of test doses, dosing frequencies, convenience for outpatients to use on their own, and blinding complexity. In some situations, an active control is also employed in a dose-ranging trial. Given the multiple treatment groups already included in a study, it may not be practical to add very many test doses into the same study.

In an early Phase II dose-ranging trial design, typically the total sample size is limited depending on budget, ethical concerns, availability of patients, and difficulty in recruitment. Under this circumstance, more treatment groups in a design imply fewer patients per treatment group. A smaller sample size of each treatment group provides a less precise estimate of treatment effect. Hence merely adding dose groups to a study design does not necessarily make the design to deliver more

informative results. A well-thought through design strategy is needed before the number of dose groups can be determined.

In our opinion, for the first-dose ranging study design, it is more important to cover a wide dose range, than simply adding more doses to cover a narrow range of doses. This suggestion is both practical and sound. In practice, a trial with four or five test doses, plus a placebo control (a total of five or six treatment groups), will deliver a good understanding of where the test medication is most active, if the dose range is wide enough and the dose spacing is reasonable. Some simulation studies (Yuan and Ting 2014) suggest that, among the number of treatment arms (4, 5, 6, or 7, including placebo), it looks like three test doses (the 4-arm design) could be insufficient in some cases. The performance increases when more than three doses are studied.

On the other hand, six test doses (the 7-arm design) may not necessarily deliver better results than the four test doses (the 5-arm design) or five test doses (the 6-arm design) comparisons. When the total sample size is fixed, the 7-arm design offers a smaller sample size per group when equal sample size is used, and the precision on the treatment effect would be sacrificed. In this case, it may be a good idea to consider an unbalanced sample size allocation, for example by allocating more subjects to highest dose and placebo arm. Therefore, from a practical point of view, a dose-ranging design including four to five test doses (in addition to placebo), which cover a wide dose range, may be very useful in designing the first-dose ranging clinical trial.

5.8 MCP-Mod

Pinheiro, Bretz and other authors introduced the concept of MCP-Mod in the first decade of the new millennium as seen in Pinheiro et al. (2006). This approach is currently widely used in the pharmaceutical industry and biotech companies. The basic idea of MCP-Mod is to consider a dose-ranging trial as a combination of two major components—the MCP part serves as the confirmatory process (hypothesis tests) and the Mod part serves as exploration (or estimation) of the dose-response relationship. This approach is very sensible in the dose-ranging trials because it helps clarify the two most important statistical ingredients in the design and analysis of a dose-ranging clinical trial: confirm and explore.

In MCP-Mod, MCP represents multiple comparison procedure, which is applied to select the dose-response model(s) for analysis. Use of MCP allows the alpha to be controlled during model selection, which serves as the confirmatory part for the study design. Mod represents modeling, which serves the purpose of estimation. In other words, after the MCP is applied to select the best model, then this selected model is used in the estimation process. One of the most important contributions of MCP-Mod is to help recognize that, although MCP has traditionally been used as a dose-selection method, it can also be used for model selection. Once this point is clarified, the flexibility of MCP applications in dose-ranging clinical trials can greatly increase.

The MCP, the first approach, takes dose as a discrete factor and generally makes few assumptions about the underlying dose-response relationship. When there is no assumed ordering imposed on the testing of multiple doses, multiplicity adjustment is necessary to protect the familywise error rate (FWER). Popular procedures such as the Bonferroni procedure, Holm procedure, and Hochberg procedure require splitting of alpha, and the level of significance (false positive rate), which results in an increased sample size.

The modeling approach (Mod), the second approach, treats dose as a continuous variable and assumes a functional relationship between the response and the dose. Sometimes the multiplicity issue is avoided as a result of the assumed functional relationship (e.g., non-decreasing dose-response relationship). The problem is potential bias caused by model misspecification.

MCP-Mod, the third (and hybrid) approach, addresses this issue by first selecting the dose-response model using MCP to protect the FWER; then the selected model is used to estimate the dose-response relationship. This approach is flexible in the sense of model selection; on the other hand, it pays the price of splitting α at the model selection step.

Although MCP-Mod was originally developed for analyzing dose-ranging clinical trials, it can also be a useful tool for study design. Details regarding the use of MCP-Mod in data analysis will be covered in Chap. 9. The focus of the current chapter is about design of dose-ranging trials, and the discussions that follow are to describe how MCP-Mod can be used in study designs.

The application of MCP-Mod in study design can be simplified to be the two-step confirm and explore process. The first step is the use of MCP to perform the confirmatory step, where alpha protection can be achieved using MCP. After the model is selected (confirmed), then this model is applied for exploration of dose-response relationships (estimation). At the design stage, the project team discusses potential models to be used for the possible dose-response relationships. After the candidate models are selected, then these models are included in the design, and MCPs are applied to control the alpha for model selection.

As a way to further simplify this procedure, a single model can be proposed (Ting 2009; Wang and Ting 2012) to implement the dose-ranging study design. Based on extensive simulations, it is apparent that a trend test can be very powerful under the monotonicity assumption. The trend test can be easily expressed as a contrast and employing such a powerful contrast test requires only a one degree of freedom and hence the entire MCP reduces to one comparison. Hence no alpha adjustment will be necessary. The statistical hypothesis can be expressed as this contrast and the entire alpha of the study is devoted into this test.

After the results demonstrate the trend test is statistically significant, the dose-response relationship can be estimated using this simple model. In fact, because the estimation of dose-response relationship does not involve any alpha consequence, such an objective can be achieved with more flexibility. It is important to understand that the purpose of a Phase II dose-ranging clinical trial is to help in describing a range of doses for the next Phase II dose-ranging exploration or for the Phase III confirmation. The ultimate objective of such a trial is not to

"confirm" minimum effective dose nor to "select" several fixed dose for Phase III. Instead, it is to help in establishing the range of doses for further exploration. Of course, based on efficacy and safety results from such a Phase II dose-ranging study, one or more doses could also be selected for confirmatory purposes and to be used in Phase III.

Given this understanding, the design thinking of "confirm, and then explore" can be implemented in such a way that a trend test is used for confirmation purposes. After a statistically significant result is obtained from this trend test, dose responses can be estimated from the observed data. One of the most important point of this strategy is the understanding that the study objective is not to identify any particular dose being a specific anchor (such as MinED, ED50 or ED90). Rather, the objective is to determine whether there is a dose-response and, after such a relationship is established, the next step is to explore or to estimate a range of doses that could potentially be confirmed in future studies. Discussion of applications of MCP, modeling, or MCP-Mod in analysis of dose-ranging trials can be found in Chap. 9.

5.9 Sample Size Considerations

Whenever sample size calculation is needed for a given study, the first point is to understand the hypothesis to be tested. In the design of a Phase II dose-ranging clinical trial, the statistical hypothesis associated with such a trial is typically a proof-of-concept hypothesis. In the case when a Phase II PoC study designed with only two treatment groups, a placebo group is compared with the MTD of study product, the hypothesis can be expressed in a straightforward fashion—responses from the test treatment is compared against responses from placebo. If the comparison is based on treatment means, then the hypothesis can be written as

$$H_0 : \mu_T = \mu_P \quad \text{versus} \quad H_0 : \mu_T > \mu_P,$$

where μ_T is the test treatment group mean and μ_P is the placebo treatment group mean.

In other words, if mean responses from the test treatment group is statistically significantly improved than the mean responses from the placebo group, then the concept is considered proven.

However, in a dose-ranging trial design, the primary hypothesis is not straightforward. In this case, the trial includes more than two treatment groups. A typical dose-ranging design consists of the highest dose of the test product (could be the MTD) on the upper end and of the zero-dose group in the form of placebo Additionally, such a design includes one or more different doses in between these two extreme treatment groups. Under this circumstance, two possible options could be considered in specifying the primary hypothesis. One option is to keep the hypothesis the same as that in the two-group setting—namely, only test the highest dose (may be MTD) against placebo.

Another option is to use a contrast which includes more than two treatment groups in order to perform the primary hypothesis test. If this second option is selected, then there could be many varieties in writing such a contrast.

For example, in a four-group design with placebo, low dose, medium dose and high dose, let μ_L be the mean response of low dose, μ_M be the mean response of the medium dose, and μ_H be the mean response of high dose. The contrast can be a trend test based on all doses (Wang and Ting 2012; Ting 2009) assuming a monotonic dose-response relationship:

$$H_0 : -3\mu_P - \mu_L + \mu_M + 3\mu_H = 0 \quad \text{versus} \quad H_0 : -3\mu_P - \mu_L + \mu_M + 3\mu_H > 0$$

Another possible proof-of-concept contrast may also include (which assumes all test doses are equally effective) is

$$H_0 : -3\mu_P + \mu_L + \mu_M + \mu_H = 0 \quad \text{versus} \quad H_0 : -3\mu_P + \mu_L + \mu_M + \mu_H > 0$$

Alternatively, it can be assumed that only the high dose is effective:

$$H_0 : -\mu_P - \mu_L - \mu_M + 3\mu_H = 0 \quad \text{versus} \quad H_0 : -\mu_P - \mu_L - \mu_M + 3\mu_H > 0$$

As can be found from these sets of hypotheses, if a contrast (which includes more than two treatment groups) is used for the statistical hypothesis, then there could be many possible ways of writing such a contrast.

Once the statistical hypothesis, the clinical difference (after adjusted for standard deviation), and the corresponding alpha (α) and beta (β) have been determined, sample size for a given dose-ranging trial can be calculated. Based on the above discussion, the sample size of a dose-ranging study can be calculated using the following formula (Chang and Chow 2006, which also appeared in Chap. 4, Sect. 4.8).

$$n = \left[\frac{(z_{1-\alpha} - z_{1-\beta})\sigma}{\delta}\right]^2 \sum_{i=0}^{k} \frac{c_i^2}{f_i}$$

where

n	is the total sample size of all treatment groups combined,
$z_{1-\alpha}$ and $z_{1-\beta}$	are the critical values of the standard normal distribution corresponding to Type I and Type II error, respectively,
δ	is the treatment difference between test product and placebo,
σ	is the standard deviation,
f_i	is the fraction of number of patients in treatment group i,
c_i	is the contrast coefficient that is associated with treatment group i.

In addition, it is assumed to have k dose groups plus placebo, where is denoted as the 0 dose group.

In the calculation of sample sizes for a dose-ranging study, two popular types of patient allocation are possible: (1) balanced allocation where each treatment group receives an equal number of patients and (2) imbalanced allocation where the highest dose and the placebo groups receives more patients and other dose groups receives fewer patients per group. The sample size formula above works well with either type of patient allocation. Examples of these types of allocation can be found in Yuan and Ting (2014). The above formula applies to cases where the primary endpoint is a continuous variable. In fact, sample size formula based on other types of response variables such as binary or time-to-event can also be found in Chang and Chow (2006).

As an illustration of how an early Phase II dose-ranging trial is designed, a real-world example is discussed in the next section. This example reflects three key features of the thinking process involved: (1) dose-range, (2) number of treatment groups, and (3) sample size. Note that in the real-world setting, other considerations are also important—selection of primary endpoint, choice of clinically meaningful difference (δ), and determination of Type I and Type II error rates, as well as statistical hypothesis to be tested. All of these considerations involve the steps on sample size calculation. Therefore, in dose-ranging study designs, once the dose-range and number of treatment groups are determined, the other design considerations are very similar to those for designing an ordinary clinical trial.

5.10 Application Example

Rofecoxib® is a medication developed by Merck for the treatment of rheumatoid arthritis (RA) and osteoarthritis (OA). In late 2004, Merck withdrew Rofecoxib® from market because of concerns with excessive cardiovascular events. After this withdrawal, the drug development environment for any new anti-inflammatory agents became more and more challenging. For every new product developed to treat RA or OA, both drug safety and efficacy are followed very closely to ensure that the efficacy of this new product can be well established, while there is no signal of any safety concerns. The clinical development plan will now involve after the PoC and early-dose ranging studies to demonstrate drug efficacy, large scale, long-term Phase III studies to follow for cardiovascular safety and gastro-intestinal (GI) safety, as well as overall safety.

A true case study of developing a compound treating osteoarthritis (OA) is introduced here as an example (Ting 2009). Consider a four-week study where OA patients first go through a screening visit. Then, after a wash-out of the current OA medication, patients are randomized to treatments into this study. Baseline measurements (demographic, medical history, background medications, efficacy, and other data) are collected prior to the first dose of study drug. Thereafter, over the four-week treatment period, efficacy and safety measurements are regularly collected at each visit. Note that at the early stage of clinical development, the length of study is constrained by the time length of animal studies completed in the

non-clinical experiments. In this case study, the animal toxicity studies are followed for up to 4 weeks and, therefore, the current clinical study is designed with a four-week study period.

In order to offer the best opportunity for the new compound to show efficacy at the early Phase II stage, the test dose to be used in the PoC should be as high as possible. In a typical clinical development program for drugs treating OA, both the efficacy dose response and the safety dose response relationships are assumed to be monotonic. If the very high dose of this new product does not demonstrate drug efficacy, the sponsor can stop developing this product and allocate investments and resources into developing other compounds with more potential to become successful drugs.

Given the requirement of demonstrating long-term drug safety, it is critical for the new compound to offer doses low enough so that in the Phase III studies where patients are exposed to long-term treatment, there is no safety concern, and patients can still experience the treatment benefit of efficacy. Hence it becomes desirable that in the PoC study, some low doses of this new compound are explored to help detect a MID. For these reasons, it is of interest to study a wide range of doses, so that the high dose that is close to MTD to help with PoC, and the low dose is low enough to help establish MinED.

In this example, the project team decides to study five parallel dose groups, which cover a wide enough dose range. In addition to these five dose groups, a placebo control and an active control are also used to help confirm the efficacy and explore the safety of the drug candidate. In most of the early Phase II clinical studies, there is no active control group. Active controls are more likely to be used in Phase III development. The reason to choose an active control in this case study is, if the results fail to distinguish between placebo and the tested doses, it will be unclear as to whether the test drug is truly ineffective or there happens to be a very high placebo response.

Under this situation, the comparison between an active control and placebo helps the project team to differentiate these two conditions. If the active control is clearly superior to placebo, then it can be concluded that the test drug is not efficacious. On the other hand, if the data cannot differentiate between the active control and the placebo, then this study is not informative because it is still unknown as to whether the test drug is beneficial or not. Hence, in this study, the active control is used for reference purposes only.

For this test drug, the MTD is estimated from results observed in previous studies. Furthermore, the formulation group has already prepared tablets with fixed drug strengths. Let the dose at MTD be 100% of the dose strength and the formulated tablets be at 2.5% and the 12.5% strengths of MTD. Given this, five doses are selected using the BDS approach. In this case study, the doses selected are placebo, 2.5, 5, 12.5, 25, and 75% of MTD, plus an active control. Based on this design, the dose range is 30 ($= 75\%/2.5\%$). The active control group is selected to be a popular NSAID (non-steroidal anti-inflammatory drug).

Although osteoarthritis is considered an inflammatory disease, the symptom that brought an OA patient to the doctor is usually joint pain. Therefore, in clinical

studies of OA, the primary measure of efficacy focuses on pain or symptoms related to pain. One of the widely accepted clinical measures for OA is the Western Ontario and McMaster University (WOMAC) Osteoarthritis index (Bellamy et al. 1997). The WOMAC score includes 24 questions with five of these questions relating to the pain domain, two relating to the stiffness domain, and 17 relating to the function domain. In this true case study, the WOMAC pain domain is selected as the primary endpoint for efficacy analysis.

From the alpha-protection point of view, it is important that each study be designed with one primary efficacy endpoint, one primary comparison, and one primary analysis set at one primary time point. Under these pre-specified conditions, it is less likely that the probability of Type I error (alpha) be artificially inflated. In this case study, the primary endpoint is change in the pain domain of the WOMAC score from baseline to week 4 (the primary time point for analysis). The primary analysis set is the intent-to-treat set; typically, this set is defined to include all patients who received at least one dose of randomized study treatment.

At this point, one of the most important considerations in alpha control is the primary comparison. Given that the design consists of five test doses plus placebo and an active control, the question is how to set the primary comparison so that the PoC decision can be made regarding this test drug. In this case study, a trend contrast with discrete doses is proposed to serve as the primary comparison. For this contrast, the active control is not included in the comparison. As discussed earlier, the NSAID to be used in this study is only for validation purposes, in case the concept is not proven (refer to Sect. 3.1 example). Hence the proposed trend contrast includes only the five dosing groups plus placebo. Here only one contrast is used and therefore no MCP adjustment. This is a simplified application of the MCP-Mod method.

For a simple trend contrast with discrete doses, the treatment groups are simply ordered from low to high. The placebo response is assumed to be 0, the lowest dose (2.5% of MTD) is assumed to have 20% response, the next lowest dose (5% of MTD) is assumed to have 40% response, the medium dose (12.5% of MTD) is assumed to have 60% of response, the second highest dose (25% of MTD) is assumed to have 80% response, and the high dose (75% of MTD) is assumed to have 100% response. Then the coefficients −5, −3, −1, 1, 3, and 5 are assigned respectively to these treatment groups coefficients of contrasts under various number of treatment groups, as given in Table 5.1. Hence the primary hypothesis to be tested is as follows:

$$H_0 : -5\mu_0 - 3\mu_1 - \mu_2 + \mu_3 + 3\mu_4 + 5\mu_5 \leq 0 \quad \text{versus}$$
$$H_A : -5\mu_0 - 3\mu_1 - \mu_2 + \mu_3 + 3\mu_4 + 5\mu_5 > 0$$

where

μ_0 denotes the mean of placebo group,
μ_1 denotes the mean of the 2.5% dose group,
μ_2 denotes the mean of the 5% dose group,

Table 5.1 Coefficients to be used in contrast for the trend test

Number of doses plus placebo	Coefficients						
	Placebo	Lowest dose	Doses increase from left to right				Highest dose
Two doses	−1	0					1
Three doses	−3	−1	1				3
Four doses	−2	−1	0	1			2
Five doses	−5	−3	−1	1	3		5
Six doses	−3	−2	−1	0	1	2	3

μ_3 denotes the mean of the 12.5% dose group,
μ_4 denotes the mean of the 25% dose group, and
μ_5 denotes the mean of the 75% dose group.

In the case of a dose-response study with k dose groups and placebo, the sample size estimates based on a contrast can be derived as follows (Chang and Chow 2006).

Let μ_i be the population mean for group i, $i = 0, \ldots, k$, where μ_0 is the mean from the placebo group and let c_i be the corresponding coefficients in the contrast. The null hypothesis of no treatment effect can be written as follows:

$$H_0 : L(\mu) = \sum_{i=0}^{k} c_i \mu_i \leq 0 \quad \text{versus} \quad H_A : L(\mu) = \sum_{i=0}^{k} c_i \mu_i = \delta > 0$$

Note that $\sum_{i=0}^{k} c_i = 0$. Under the alternative hypothesis and the condition of homogeneous variance, the sample size per group can be obtained as $n = \left[\frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\delta}\right]^2 \sum_{i=0}^{k} c_i^2$, where $z_{1-\alpha}$ and $z_{1-\beta}$ are the corresponding quantile points of a standard normal distribution.

Based on published data, the standard deviation is about 4 ($\sigma = 4$). From the observed placebo response, and the anticipated response from this test drug, possible targeted clinical differences can be considered including 1.00, 1.25, or 1.50. From such a large standard deviation, and a relatively small delta, the required sample size could be very large. For example, if a pairwise comparison is used, then under one-sided alpha of 0.05, with 90% power and delta of 1.25, the sample size for each treatment group would be approximately 176 subjects. For a 7-arm study, this criterion will take a total sample size of 1232 subjects. For a study powered using the trend contrast as specified above, the sample size would reduce to 126 per group (882 overall, across the 7 groups).

Suppose the study with the same assumptions is powered to detect a significant trend contrast test under a balanced design with equal number of subjects in each treatment group. In Table 5.2, a scenario is considered for different sets of alpha (α), power (β), and delta (δ). The standard deviation is assumed to be 4, and the

Table 5.2 Sample sizes under various one-sided alpha, statistical power, and delta (difference in means)

Alpha	Power	Delta	Sample per group	Total sample
0.10	0.9	1.00	151	1057
0.05	0.8	1.00	142	994
0.05	0.9	1.00	196	1372
0.10	**0.9**	**1.25**	**97**	**679**
0.05	0.8	1.25	91	637
0.05	0.9	1.25	126	882
0.10	0.9	1.50	67	469
0.05	0.8	1.50	63	441
0.05	0.9	1.50	87	609

Note Calculations assume a standard deviation (σ) of 4. Contrast used to perform this sample size estimation is $-5\ \mu_{0-3}\ \mu_1 - \mu_2 + \mu_3 + 3\ \mu_4 + 5\ \mu_5$

response is assumed to be 1 unit, 1.25 units, or 1.5 units. The sample sizes under various alpha, power and delta are presented in Table 5.2.

Since the test drug is at an early stage of clinical development, and the concept has not been proven, it is difficult to justify a huge budget of 882 subjects in an early Phase II study. Upper management discussed the matter with the project team and decided that they are willing to take a higher risk on alpha and reduce the sample size to meet the budget requirement. The compromise is that the two-sided alpha can be increased to 0.2 (or one-sided alpha of 0.1) and keep the power at 90%.

For this study, a treatment difference from placebo of approximately 1.25 was selected. From Table 5.2, a sample size of 97 per treatment group provides estimated power of 90% for an one-sided alpha of 0.10 and, for this sample size, the power is over 80% for one-sided alpha of 0.05. A sample size of 100 subjects per treatment group is sufficient to detect differences larger than 1.25 in the mean change from baseline in the WOMAC OA Index Pain Subscale, using the aforementioned linear contrast. Since a single test statistic for trend is employed, no adjustment for multiple comparison is needed. If this statistical test is significant at alpha of 0.1 (one-sided) MID will be estimated.

5.11 Discussion

There are many challenges in designing the first dose-ranging clinical trial. Two main problems surface. First, there is not much of information about the efficacy of the candidate product available. Second, too many answers are expected for this trial to deliver. Project team members tend to anticipate such a study will establish the PoC, identify the MinED, characterize the dose-response relationship, and recommend doses for Phase III designs. Fundamental statistical training points out

that the best design is that one study answers one question. Hence the nature of a dose-ranging trial makes it a very difficult design in order to address many scientific questions and with sparse information available before the study design stage.

In fact, there should be a logic sequence of thinking in addressing these relevant questions. The first step is to consider how wide a dose range should be—which is the most critical step. As a way to approach this question, information about the candidate under development, other marketed drugs in similar class, the indications this study is designed for, the target product profile, and characteristics of the primary endpoint, as well as other considerations, can all contribute to help with a better understanding of the potential dose range need to be explored. In practice, it is also important to have a good understanding of dosing frequency (from Phase I results), formulation (from manufacturing), and other relevant issues.

After the dose range is determined, the next step would be to consider the number of treatment groups. Will there be a need of an active control? How many doses should be included in this design? What dose spacing should be considered? Basically speaking, a trial with four or five test doses, plus a placebo control, will deliver a good understanding of where the test medication is most active, if the dose range is wide enough and the dose spacing is reasonable. Extensive simulations indicate that six or more test doses (in addition to placebo) design may not necessarily deliver better results than the four test doses or five test doses (Yuan and Ting 2014). A study design with too many doses may introduce additional complexities from practical and logistical points of view (e.g., formulation, blinding).

Sample size calculation is needed for almost every clinical trial. Considerations in sample size calculations for a dose-ranging trial are very similar to those for other trials. However, because the nature of a dose-ranging trial includes multiple treatment groups, this feature may cause some level of confusion both to statisticians and to non-statisticians. Nonetheless, if the decision is to only use a single degree of freedom contrast test as the primary comparison, then the situation can be simplified and addressed successfully. In the case of one contrast, the two major challenges would be what assumptions of dose-response relationship the team is willing to make and sample size allocation, either a balanced approach (an equal number of patients per group) or an imbalanced approach (an unequal number of patients per group). After these decisions are made, the statistician would then follow up and calculate the sample size for the study.

Suppose a contrast with one degree of freedom is selected and the sample size allocation is determined, the typical sample size considerations such as choice of treatment difference, choice of alpha and beta can then be specified. With these information, sample size can be calculated using existing formula. An example of a real study design in a test product developed for treatment of osteoarthritis is included in this chapter to demonstrate how all of these key elements thinking can be applied in a real-world situation.

References

Bellamy, N., Campbell, J., Stevens, J., Pilch, L., Stewart, C., & Mahmood, Z. (1997). Validation study of a computerized version of the Western Ontario and McMaster Universities VA 3.0 osteoarthritis index. *Journal of Rheumatology, 24,* 2413–2415.

Chang, M., & Chow, S. C. (2006). Power and sample size for dose response studies. In *Dose finding in drug development* (pp. 220–241). New York: Springer.

Farrar, J. T., Young, J. P., LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain, 94,* 149–158.

Filloon, T. G. (1995). Estimating the minimum therapeutically effective dose of a compound via regression modelling and percentile estimation. *Statistics in Medicine, 14,* 925–932.

Hamlett, A., Ting, N., Hanumara, C., & Finman, J. S. (2002). Dose spacing in early dose response clinical trial designs. *Drug Information Journal, 36*(4), 855–864.

Hochberg, Y., & Tamhane, A. (1987). *Multiple comparison procedures*. New York: Wiley.

Katz, N. P., Paillard, F. C., & Ekman, E. (2015). Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions. *Journal of Orthopaedic Surgery and Research, 10,* 24.

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

McLeod, L. D., Cappelleri, J. C., & Hays, R. D. (2016). Best (but oft-forgotten) practices: Expressing and interpreting associations and effect sizes in clinical outcome assessments. *American Journal of Clinical Nutrition*.

Osoba, D., Rodrigues, G., Myles, J., Zee, B., & Poter, J. (1998). Interpreting the significance of changes in health-related quality-of-life scores. *Journal of Clinical Oncology, 16*(1), 139–144.

Pinheiro, J. C., Bretz, F., & Branson, M. (2006). Analysis of dose-response studies—Modeling approaches. In *Dose finding in drug development* (pp. 146–171). New York: Springer.

Rosen, R. C., Allen, K. R., Ni, X., & Aruujo, A. B. (2011). Minimal clinically important differences in the erectile function domain of the international index of erectile function scale. *European Urology, 60*, 1010–1016.

Thomas, N., Ting, N. (2009). Minimum effective dose. In *Encyclopedia of clinical trials.* Hoboken, NJ: Wiley-Blackwell.

Ting, N. (2008). Confirm and explore, a stepwise approach to clinical trial design. *Drug Information Journal, 42*(6), 545–554.

Ting, N. (2009). Practical and statistical considerations in designing an early phase II osteoarthritis clinical trial: A case study. In *Communications in statistics—Theory and methods* (Vol. 38(18), pp. 3282–3296).

Wang, X., & Ting, N. (2012). A proof-of-concept clinical trial design combined with dose-ranging exploration. *Biopharmaceutical Statistics*, wileyonlinelibrary.com. doi:10.1002/pst.1525

Yuan, G., & Ting, N. (2014). First dose ranging clinical trial design—More doses? Or a wider range? In *Clinical trial biostatistics and biopharmaceutical applications*. Taylor & Francis.

Chapter 6
Combining Proof of Concept and Dose-Ranging Trials

6.1 Background

Chapters 4 and 5 discuss proof of concept (PoC) and dose-ranging studies respectively and separately. However, sometimes it is desirable to combine PoC and dose-ranging studies into one single clinical trial. The advantage of such a design is to first make a Go/NoGo decision based on PoC. If the decision is "Go," then the same study would subsequently provide dose-ranging information to help design the next study. When well designed and analyzed, the combined study saves development time by providing a range of efficacious doses going forward. The disadvantage is more investment would be needed before the concept is proven or not proven. In case the test drug does not work, this translates into larger sunken and opportunity costs. Again, as mentioned in Chap. 5, in designing dose ranging trials four to five doses (Yuan and Ting 2014) plus placebo would be appropriate and the use of binary dose spacing (Hamlett et al. 2002) is highly recommended.

The idea of combining a PoC and a dose-ranging study is very similar to the consideration of confirm and explore, as discussed in Chap. 3. The PoC objective involves a confirmatory practice—this can be achieved by performing statistical hypothesis tests. Only after the concept is proven, the natural next step will be dose-range exploration. On the other hand, if the concept is not proven, then the conclusion is that the drug candidate cannot deliver the efficacy as expected; therefore, there is no need to explore the dose-response relationship. The PoC hypothesis needs to be simple and clear. Ideally, this would be a single degree of freedom test, without any multiplicity adjustment of alpha. In a one degree of freedom, simple hypothesis test setting, the PoC decision should be relatively easier and more straight forward.

However, in a combined PoC and dose-ranging study, there are multiple treatment groups—a placebo control plus several doses of the test drug. Hence under this setting, some statisticians tend to perform multiple comparison procedure (MCP) adjustment from the hypothesis test point of view. The difficulty is that once

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_6

MCP is performed, PoC becomes a complex question—it is no longer a single degree of freedom hypothesis testing to answer a Go/NoGo question. The key to avoid this confusion is to separate PoC from dose-ranging and see them as two different questions. These two questions are addressed sequentially—confirm PoC first and then explore the dose-response relationship.

6.2 Considerations in Designing Combined PoC and Dose Ranging Studies

In a two-group PoC study, the confirmatory step is relatively simple: compare the test drug [typically at a high dose or the maximum tolerated dose (MTD)] against the placebo control. But this step could be confusing in a PoC combined with dose-ranging clinical trial design, even when the PoC step is simplified into a single degree of freedom test. In other words, what does it mean that the concept is proven? In fact, the simplest approach can still be that the comparison of highest dose against placebo—by ignoring the low or medium doses. As discussed in Chap. 4, this is based on the monotonicity assumption.

Another option is to combine all test doses and compare this combination against placebo. This approach is not recommended if any two doses could provide different efficacy responses. As a result of averaging all test doses, the dose with lower efficacy response could drag the combined dose effect to be closer to placebo. More on this point will be discussed in later sections. A different approach could be use of MCP to select at least one effective dose. For example, if Bonferroni procedure is used, then if any dose is significantly different from placebo using the Bonferroni adjustment, then the concept is proven. Again, use of MCP is not recommended for PoC because these MCP approaches are generally less powerful.

One popular approach is to propose a single degree of freedom contrast test. A contrast can be very flexible, and any possible linear combination of doses can be expressed in a contrast comparison. Among all potential contrasts, Wang and Ting (2012) recommend the use of a trend test. The authors performed large amount of simulations and, generally speaking, if the monotonic efficacy dose-response relationship can be assumed, then the trend test is one of the most powerful ways of evaluating PoC.

One question of using a trend test would be what if the underlying dose-response relationship is not monotonic? For example, what to do if the drug under development is for a CNS indication (where monotonicity is not necessarily assured)? Under this circumstance, our recommendation is to use a two-group comparison of highest dose against placebo to establish PoC. However, a trend test can still be considered in developing a CNS drug. What follows are some arguments to support this proposal.

First, even though the population dose-response is not monotonic, the individual dose-response could still be monotonic. Second, for the drug under development,

the part of non-monotonicity may be beyond MTD. In other words, for the given range of doses to be studied (doses below MTD), the dose-response relationship could still be monotonic. Third, based on simulations, as long as the violation of monotonicity is not too strong, a trend test is still relatively powerful.

The trend test is implemented using a one degree of freedom contrast, where the coefficients of this contrast mimic those coefficients of a linear contrast. For a given dose-ranging study, coefficients for the proposed trend test can be found in Table 5.1 (Wang and Ting 2012). In this table, doses of the study drug are arranged in an ascending order; that is, the first coefficient always corresponds to the placebo treatment group, the last coefficient corresponds to the highest dose group, and other coefficients correspond to the doses arranged in an ascending order. In practice, there is rarely any fixed dose-ranging design employing more than six test doses; therefore, Table 5.1 provides these coefficients up to six test doses.

For example, in a design with five doses (see example in Sect. 5.10), the null and alternative one-sided statistical hypotheses are

$$H_0: -5\mu_0 - 3\mu_1 - \mu_2 + \mu_3 + 3\mu_4 + 5\mu_5 \leq 0 \quad \text{versus}$$
$$H_1: -5\mu_0 - 3\mu_1 - \mu_2 + \mu_3 + 3\mu_4 + 5\mu_5 > 0,$$

assuming that higher responses are more favorable.

If the one-sided trend test is not significant at level the stated level of significance (alpha, α) then claim the concept is not proven and further development of the study drug is stopped. On the other hand, if the trend test is statistically significant, then the concept is proven and a "Go" decision is made about the drug candidate. The proposed method is to estimate dose-response relationship using the doses included in the study design.

In a combined PoC and dose-ranging clinical trial, the recommended PoC hypothesis test is either using the highest dose to compare against the placebo control or performing a trend test using a trend test contrast. The most important understanding is that, in such a combined study, the first step should be PoC and, only after PoC, then estimate dose-response relationship. It is also of interest to note that dose-ranging, by nature of the problem, is an estimation practice. No need for hypothesis tests. This point is further elaborated in later sections of this chapter.

6.3 Concerns of Using a Dose-Response Model

It is common that a dose-response model is used in analyzing dose-ranging trials. However, from a study design point of view, the use of such a model is typically based on additional assumptions. It is well known that "all models are wrong, some are useful" (Box 1987). Designing dose-ranging trials based on a wrong model, or wrong parameters (even when the underlying model is correct, the parameter values may be wrong) could potentially lead to undesirable consequences. In data analysis,

when blind is broken, efficacy dose-response data are observed. At this point, use of a model that "fits the data well" is a reasonable thing to do. But it could be risky to use a model at the design stage.

As indicated earlier, the two key assumptions necessary in designing a dose-ranging trial are the correct MTD and monotonicity in efficacy. Based on the experiences of the authors, any additional assumptions about the shape of the underlying dose-response relationship could be potentially misleading. Hence, although modeling is a useful tool for data analysis, it is not recommended to use models at the study design stage unless information is available to make reasonable assumptions for a dose-response model.

In fact, when a model is used in designing a dose-ranging trial, the underlying assumption could be relatively strong—not only does that the selected model have to be right but also its parameters have to be all correct. More importantly, the model has to assume the explored range of doses being correct. In practice, these assumptions could be incorrect, especially when the range of active doses guessed is wrong. Many very expensive mistakes are made in practice because the designed dose-range was too narrow or that the dose-response model was mis-specified. Over the years, experiences indicate that these two critical assumptions—correct MTD and monotonicity—are necessary for study design. In some practical situations, even these two simple assumptions may not be met. Any additional assumption could potentially lead to very ineffective study design and expensive failures. At the time of designing a trial, if there are too many unknowns, adding assumptions would add potential risks to the design. In the advancement of science, expensive mistakes are often made not because "we didn't know," but because "we thought we knew." The real-world cases repeatedly demonstrated that we thought we knew the model, and studies were designed according to the model, which eventually led to wasteful failures. One example can be found in Sect. 5.3, other examples include approved drugs whose labeled (and approved) dose needs to be increased or decreased after during the post-approval phase.

Given this background, the authors would warn readers to check every additional assumption very carefully by assessing factually the robustness of all the available information, such that any further assumptions supporting the model is valid before it is used in designing the dose-ranging clinical trial. For this same reason, the discussion regarding specifically study design in this chapter is basically model-free with the intent to separate somehow the thinking during study design stage from the analysis stage. Modeling can be valuable for analyzing the data to better understand the dose-response relationship and to provide a better dose-regimen selection. It still remains important to have proper consideration as to whether the method used for sample size estimation is more or less conservative than the method to be used for data analysis.

6.4 Sample Size Allocation

In a PoC combined with dose-ranging design, the sample size calculation would naturally be based on the PoC hypothesis test. However, there may still be an interest in comparing each dose against placebo. Under this setting, one of the more powerful ways of making multiple comparison adjustment would be the gate-keeping multiple comparison procedure (MCP). The gate-keeping procedure (Dmitrienko et al. 2003) is to start testing with the highest dose against placebo, using the entire study-wise alpha. If the highest dose is statistically significant at alpha, then test the next highest dose against placebo with the entire alpha. If the highest dose is not statistically significant, stop and claim none of the doses is different from placebo. Repeat this testing procedure until either a particular dose is not significant (in this case, stop testing and claim all doses higher than this dose are significant, and this dose together with all lower doses are not significant) or all study doses are tested. Gate-keeping MCP is very powerful and widely accepted in the pharmaceutical industry; however, it strongly depends on the monotonicity assumption.

We propose two ways of writing PoC contrasts: (1) the traditional two-group comparison (i.e., highest studied dose against placebo) and (2) a trend test. If the decision is to follow the traditional two-group contrast to test the PoC hypothesis, then the sample size can be calculated using two-group contrasts, without multiple comparison adjustment. If the concept is proven, then follow the gate-keeping procedure to test from highest dose downward to lower doses, each against placebo. Such a strategy begets several advantages, including: Type I error rate (α) is well protected, the procedure is simple to understand and to implement; it can be easily explained to non-statisticians; mistakes are less likely to occur in real-world applications under monotonicity assumption; and it is among the most powerful ways of testing pairwise hypotheses.

The only limitation of this strategy is its heavy dependence on the monotonicity assumption. One option of this approach is to consider not to test subsequent pairwise hypothesis. In other words, test only one PoC hypothesis and, after the concept is proven, move on to estimate a dose-response relationship. Hence no alpha will be spent after PoC. On this basis, an unequal allocation of sample size could potentially make the entire study more efficient. In a four-treatment-group design (placebo plus three test doses), an allocation of 2:1:1:2 of sample size could be powerful, efficient, and informative. Similarly, in a five-treatment-group design (placebo plus four test doses), a 3:2:2:2:3 allocation or a 2:1:1:1:2 allocation could also be considered.

The second strategy is to write the PoC hypothesis as a trend test. In other words, if the trend test is significant at level alpha, then we claim that the concept is proven and make a “Go” decision for this drug candidate. Once the concept is proven, no more hypothesis test is performed and the dose-response questions are turned into an estimation procedure. Thus all alphas are spent on the PoC, which leaves no more alpha for further pairwise comparisons. With this strategy, both equal and

unequal sample size allocations can be contemplated. Use of trend test to determine PoC may reduce the total sample size if no pairwise comparison is performed. However, if there is still an interest in testing each dose against placebo after the concept is proven, then the sample size should be calculated using the first strategy (two-group traditional PoC).

Examples of combined PoC and dose-ranging designs can be found in Ting (2009) and Wang and Ting (2012). In these articles, the proposed PoC can be achieved using a trend test. For example, in a four-group design with placebo, low dose, medium dose and high dose, the traditional PoC can be only based on high dose versus placebo (let μ_L be the mean response of low dose, μ_M be the mean response for the medium dose, and μ_H be the mean response of high dose).

Table 5.1 lists the coefficients for trend tests under a variety number of treatment groups. Because a four group (three doses plus placebo) is a popular design, a set of contrasts based on four groups can be described as

$$H_0: \mu_H = \mu_P \quad \text{versus} \quad H_1: \mu_H > \mu_P$$

or a contrast can be a trend test which is based on all doses (Wang and Ting 2012; Ting 2009), assuming monotonic dose-response relationship. The corresponding contrast for a four group design is

$$H_0: -3\mu_P - \mu_L + \mu_M + 3\mu_H = 0 \quad \text{versus} \quad H_1: -3\mu_P - \mu_L + \mu_M + 3\mu_H > 0$$

or, if only the high dose is assumed effective,

$$H_0: -\mu_P - \mu_L - \mu_M + 3\mu_H = 0 \quad \text{versus} \quad H_1: -\mu_P - \mu_L - \mu_M + 3\mu_H > 0.$$

Alternatively, if the high dose and median dose are assumed equally effective, while low dose is like placebo,

$$H_0: -\mu_P - \mu_L + \mu_M + \mu_H = 0 \quad \text{versus} \quad H_1: -\mu_P - \mu_L + \mu_M + \mu_H > 0.$$

Other possible PoC contrast may also include (assuming all doses are equally effective)

$$H_0: -3\mu_P + \mu_L + \mu_M + \mu_H = 0 \quad \text{versus} \quad H_1: -3\mu_P + \mu_L + \mu_M + \mu_H > 0.$$

As can be found from the these examples, if a contrast (that includes more than two treatment groups) is used for the PoC hypothesis, then there could be many possible ways of writing such a contrast.

6.4.1 Comparison of Power

In this section, we compare the power of five tests discussed above in Sect. 6.3. Chang and Chow (2006) indicated that for a dose-ranging study with k arms., the proof-of-concept can be tested using the following contrast test (either strategy 1 with two-group traditional PoC or strategy 2 with trend test):

$$H_0: L(\mu) = \sum_{i=0}^{k} c_i \mu_i = 0 \quad H_a: L(\mu) = \sum_{i=0}^{k} c_i \mu_i = \varepsilon$$

where $\sum_{i=0}^{k} c_i = 0$, where μ_i is the population mean for group i.

The power of the test is

$$1 - \beta = \Phi\left(\frac{\in}{\sigma}\sqrt{\frac{n}{\sum_{i=0}^{k} c_i^2 / f_i}}\right)$$

where Φ is CDF of a standard normal distribution, σ is the population standard deviation, n is the total sample size of the study and f_i is the sample size fraction for the ith group. For example, for a study with n = 60 and $f = (1/3, 1/6, 1/6, 1/3)$, 20 subjects would be allocated to the first group and the fourth group, while 10 subjects would be allocated to the second group and the third group respectively.

Given the flexibility of selecting PoC criteria, a variety of sample size allocations can be considered such as 2:1:1:2 and 3:2:2:3, which would to allow the placebo and the high dose to have more patients, with fewer patients assigned to doses in between. Other proposals can also be candidates for sample size allocations.

Under a fixed total sample size, with the aforementioned five PoC contrasts, comparisons of methods were made with equal or unequal sample size allocation. In these comparisons, the metric used is power of PoC test. Because power comparisons can be evaluated analytically, there is no need to perform simulations in these comparisons. Consider the statistical power for a study with a total sample size of 60 patients allocated to four groups: high (30 mg), median (10 mg), low dose (3 mg) of test drug and placebo. These comparisons are performed under five scenarios with different dose response relationships as illustrated in Table 6.1 and Fig. 6.1.

Table 6.1 Mean response for the four arms assuming a common standard deviation of 1

No.	Shape	μ_0 (placebo)	μ_1 (low)	μ_2 (median)	μ_3 (high)
1	Linear	0.15	0.24	0.45	1.05
2	Step	0.15	0.6	0.6	1.05
3	Quadratic	0.15	0.6	1.05	0.9
4	Convex (curving out)	0.15	0.15	0.15	0.9
5	Concave (curving in)	0.15	0.9	0.9	0.9

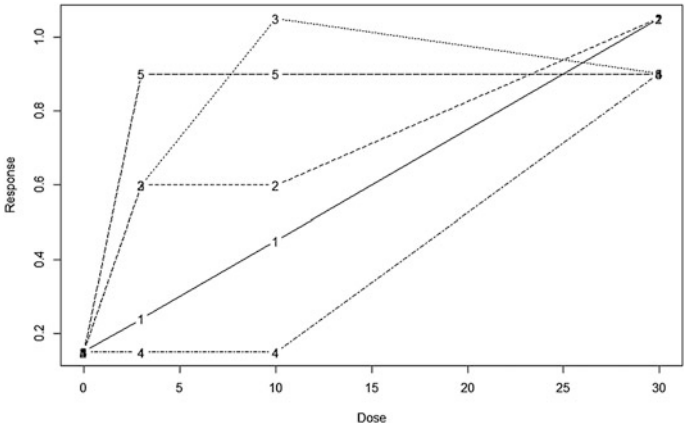

Fig. 6.1 Shapes of dose-response relationship evaluated under a four-arm design

Table 6.2 shows the statistical power of PoC given a total sample size of 60 patients, assuming the five different dose response shapes as given in Table 6.1 and Fig. 6.1. In Table 6.2, A, B, C, D, E represent the five different contrast as shown in Table 6.1. The power for total sample size of 60 is also shown as in Table 6.2. The methods with 2:1:1:2 (20:10:10:20) allocation in general has a better performance (except for B where it is slightly worse but comparable) than 1:1:1:1 (15:15:15:15) allocation. This result is not surprising given the fact that more subjects are allocated on the higher dose and the placebo, where the treatment effect can be most easily differentiated.

The most powerful yet robust method is the trend test and the traditional PoC test, which uses only information from the highest dose and placebo. Both contrasts provided more than 80% power consistently across different dose response shape with 2:1:1:2 allocation and more than 75% of power with 1:1:1:1 allocation. Although the traditional PoC test did not incorporate all data in the study, the power is comparable to trend test in most situations.

While it is not intuitive at first glance, using average of high dose and median against the average of low dose and placebo should be avoided. It only provides around 50% power under a commonly seen concave response curve, and it is almost uniformly less powerful than the trend test or the traditional PoC test (except for the umbrella shape under equal allocation where it is close to trend test). The loss of statistical power is substantial under many situations.

Using high dose against average of the other three groups should be avoided by all means, because the power will be below 40% under the concave (curving in) curve. Using the average of three doses from test drug against placebo could be an option, if the umbrella shape of curve is a real possibility, for example in the

Table 6.2 Statistical power for a trial with 60 patients in total, one-sided alpha = 0.1. Simulation results

	Method	Linear	Step	Quadratic	Convex	Concave
1:1:1:1	A: High versus PBO (−1, 0, 0, 1)	0.88	0.88	0.78	0.78	0.78
	B: Trend test (−3, −1, 1, 3)	0.89	0.85	0.85	0.75	0.75
	C: High versus median/low/PBO (−1, −1, −1, 3)	0.90	0.77	0.39	0.89	0.33
	D: High/median versus low/PBO (−1, −1, 1, 1)	0.81	0.68	0.85	0.57	0.57
	E: High/median/low versus PBO (−3, 1, 1, 1)	0.56	0.77	0.86	0.33	0.89
2:1:1:2	A: High versus PBO (−1, 0, 0, 1)	0.94	0.94	0.86	0.86	0.86
	B: Trend test (−3, −1, 1, 3)	0.93	0.90	0.90	0.81	0.81
	C: High versus median/low/PBO (−1, −1, −1, 3)	0.93	0.81	0.42	0.92	0.35
	D: High/median versus low/PBO (−1, −1, 1, 1)	0.77	0.64	0.82	0.53	0.53
	E: High/median/low versus PBO (−3, 1, 1, 1)	0.60	0.81	0.89	0.35	0.92

development of certain anti-psychotic agents. However, stemming from the significant loss of power of between 30 and 35% in case of a convex (curving out) shape for the dose-response curve, it should only be used when there is strong confidence that the possibility of convex shape can be excluded, which may not happen often in the first dose ranging study.

In fact, there are several advantages of using a trend test, two of which are stated as follows:

1. it is simple to implement; there is no need to be concern about MCP adjustment (the trend test is only a one degree of freedom test);
2. under the monotonicity assumption, a trend test is robust with respect to dose-ranges selected within a given study.

6.5 Estimation of Dose-Response Relationship

It is critical to keep in mind that the two major objectives of Phase II are PoC and recommendation of a range of doses for Phase III development. Note that some investigators consider objectives of Phase II to also include establishing the precise minimum effective dose (MinED) (ICH-E4; Thomas and Ting 2009); characterizing the dose-response relationships (or describing the entire dose-response curves); finding the optimal doses; and understanding the benefit/risk ratio. In fact, many of these additional objectives are neither realistic nor desirable at Phase II.

Within the pharmaceutical and bio-tech industries, drugs/biologics are discovered and developed for the general patient population. Hence the dose-selection process is to recommend one or a few doses for the overall patient population use, not for individual patient dosing. If the target dose is for the entire patient population, then the chosen dose cannot be optimal to each and every individual patient. Therefore, it is not practical to search for an "optimal dose" because an optimal dose is more meaningful to a given individual patient, rather than to the entire patient population.

Although the dose-response curve from Fig. 6.1 looks like continuous curves, which are only theoretical curves. In reality, drugs are formulated with discrete dose strengths. In the treatment of chronic diseases, the number of pre-formulated doses is somewhat limited. From this viewpoint, instead of "characterizing the dose-response relationships," early Phase II studies are designed to help locate the target efficacy dose range (the range of doses where the ascending part of the efficacy dose-response curve is the steepest). Therefore, it is reasonable to consider dose-finding practices as operated under a few fixed-dose levels, instead of a continuous dose-response curve.

From this point of view, MinED is not a specific number but an efficacious dose based on the existing formulation, a dose strength lower than this dose is not anticipated to deliver meaningful population efficacy response. For instance, suppose the test drug is formulated as 10 mg tablets and 20 mg (2 tablets) were shown to be efficacious, while 10 mg failed to distinguish its efficacy from placebo. Then the MinED is, practically, 20 mg. In case a model was applied to the dose-response study, and this model indicates that the MinED is 16.32 mg. In practice, such a MinED is not very useful. From the drug development point of view, it is more practical to consider 20 mg as the MinED. Hence a precise point estimate of MinED is not necessarily useful or necessary.

In most of the drug development programs, benefit/risk ratio is studied after Phase III study results are ready. The benefit-risk discussion is typically covered in the Clinical Overview document in a regulatory submission based on the review of the entire clinical database. Therefore to seek for clear benefit-risk relationship at Phase II would generally be pre-mature.

One very common misunderstanding in the Phase II clinical development process is that, when the project team planned for a combined PoC and dose-ranging trial, the team expects to move to Phase III after results are ready from this

combined trial. In general, though, it may be more efficient to establish a good understanding of efficacy and safety of various doses at Phase II, before rushing to Phase III. Traditionally, a good Phase II plan takes at least three sequential studies, before progressing a candidate into Phase III. The first is a PoC trial; the second is a dose-ranging trial capturing a wide range of doses; and the third is a dose-ranging study with a narrower range of doses focusing on the steepest ascending part of the dose-response curve.

For instance, if the second trial is designed with placebo, 1, 10, 100 and 1000 mg doses, and based on the study results, the appropriate dose range is between 10 and 100 mg. In this case the third clinical trial will be designed to study the doses around this range. After a good understanding about efficacy and safety of all the studied doses, then the drug candidate can be progressed into Phase III development. In the case when a PoC is combined with the first dose-ranging study, and if the decision is to "Go" forward and further develop this drug candidate, then at least one second dose-ranging study should be considered to "zoom in" or target the appropriate range of doses.

Dose finding is, by its nature, an exploratory practice. In many team discussions, researchers emphasize that "a decision is needed for selecting doses to move forward." Such an affirmation could easily be misunderstood as a confirmatory practice. In fact, it is not! The entire Phase II process is to propose a range of doses for contemplation of designing Phase III clinical trials. Therefore, the statistical procedure in supporting this objective is estimation, not statistical hypothesis testing. Phase II clinical data provide information for point estimates and confidence intervals about efficacy responses at various doses being studied. These estimates can then be used to help understanding the doses to be proposed for Phase III.

Note that in recommending doses for Phase III study designs, the considerations are not only restricted to efficacy. In addition to efficacy, selection of Phase III doses including considerations from safety, pharmacokinetics, and all available clinical data. One major concern in evaluating these data and selecting Phase III doses is how likely the Phase I/II findings can be extrapolated into Phase III. Experiences in drug development indicate that Phase III usually dilute the treatment effects observed from previous phases. It is also important to consider how to manage potential safety findings from patients after long-term exposure of the drug candidate. If feasible, it is advised that more than one dose of study medication be included for evaluation in Phase III clinical trials.

One very important point that project team members, as well as upper management, should keep in mind is that dose ranging is a Phase II practice; it will be too late if doses outside of the Phase II recommended range are tested at Phase III or IV. One common mistake made by almost all sponsors, and across most of therapeutic areas, is taking shortcut in Phase II clinical development. Then at the Phase III or Phase IV stages, sponsors were forced to repeat dose-finding trials at doses outside of range being studied in Phase II. Typically the cost of such a practice is very high.

Phase II is the most important phase of clinical development of new medicinal products, because the experiment units in Phase I studies generally do not include

patients with the targeted disease (and hence drug efficacy is first assessed in Phase II). The development investment for Phase III is quite expensive. Experiences indicate that careful evaluations of Phase II results before rushing into Phase III almost always lead to a better quality decision. It is critical to realize that Phase II provides the last opportunity to stop developing an unsuccessful drug candidate before committing huge amount of Phase III investments.

6.6 Risk of Inconclusiveness

In the new drug development process, the management within a sponsor company attempts to make decisions based on the study results from each of the clinical trials. Unfortunately, in many situations, the study results are inconclusive. If the results from a clinical trial cannot clearly address this question, then some studies

Table 6.3 Examples of statistically-related issues that may cause inconclusiveness from clinical studies

1	Study was underpowered because of an underestimation of variance
2	Study was underpowered because of an overestimation of effect size
3	Study was adequately powered, with results demonstrating statistical significance, but not for clinical significance
4	Only one of the co-primary endpoints reached significance
5	Study demonstrated efficacy only at the lowest dose, but no efficacy as the medium or high doses
6	Study failed on the primary endpoint but demonstrated strong trend in all the secondary endpoints
7	Study failed the primary comparison but demonstrated strong trend in a particular subgroup
8	The interim analysis showed promising trend, but the final analysis failed to demonstrate significance
9	No clear signal at interim, study continued but showed good response in the final analysis
10	Significant treatment-by-region interaction in a multi-regional trial
11	Significant results from the per protocol analysis set but not from intention-to-treat
12	Significant at a different time point other than the primary time point
13	Outliers (conclusions changed if outliers are excluded)
14	Missing covariate data where the covariate is used in the primary analysis model
15	Inconsistent conclusion when the center-pooling strategies are different
16	Inconsistent conclusion when visit windows are defined differently
17	Different analytical methods (e.g., Cochrane-Mantel-Haenzel versus logistic regression) le to different conclusions
18	Different analytical model (different covariatesor interactions in the analysis of covariance model) led to different conclusions
19	Different multiple comparison procedures led to different conclusions
20	Different data transformation (e.g., log and rank) led to different conclusions

Table 6.4 Examples of non-statistically related issues that may cause inconclusiveness from clinical studies

1	Study demonstrate both statistical and clinical significance in efficacy, but there were safety concerns
2	Placebo response was favorably too high (higher than published results)
3	In a study including test drug, placebo and active control, results failed to demonstrate that the active control was superior to placebo
4	In a non-inferiority comparison against active control, study demonstrated non-inferiority, but response rates of both test drug and active control were lower than expected or the historical control
5	Either study drug or active control reached the drug expiration date before study completion
6	Patients recruited in the study all followed the entry criteria, but the patient population in the study did not represent the underlying patient population of interest (e.g., disease too severe, disease too mild, gender imbalance)
7	Patient non-compliance
8	Imbalance in patient non-compliance
9	Active control withdrawn from the market during the study (e.g., Vioxx in 2004)
10	Dropout rates too high
11	Dropout rates imbalance across treatment groups

may have to be repeated and additional resources will need to be invested. Oftentimes these additional investments are wasted as the entire project gets terminated and, in some cases, the entire development schedule is severely delayed. Therefore the cost of inconclusiveness can be very high. Some examples of inconclusive study results are listed in Tables 6.3 and 6.4 (Ting 2011). Table 6.3 provides some statistical issues while Table 6.4 some non-statistical issues that potentially cause inconclusiveness.

In addition to those issues listed in Tables 6.3 and 6.4, data cleaning is always a difficult operational challenge. Occasionally, data errors are detected during the report review process. In some situations, the correction of some data errors may cause the change of clinical conclusions. Thus, it is crucial for the project team to design studies not only to control the Type I and II errors but also to reduce the likelihood of inconclusiveness. In fact, there are many other types of risks in inconclusiveness. This will be the topic of the next chapter.

References

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces* (p. 424). New York: Wiley. ISBN 047180339.

Chang, M., & Chow, S. C. (2006) Power and sample size for dose response studies. In *Dose finding in drug development* (pp. 220–241). New York: Springer.

Dmitrienko, A., Offen, W., & Westfall, P. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine, 22,* 2387–2400.

Hamlett, A., Ting, N., Hanumara, C., & Finman, J. S. (2002). Dose spacing in early dose response clinical trial designs. *Drug Information Journal, 36*(4), 855–864.

Thomas, N., & Ting, N. (2009). Minimum effective dose. In *Encyclopedia of clinical trials*. Hoboken, NJ: Wiley-Blackwell.

Ting, N. (2009). Practical and statistical considerations in designing an early phase II osteoarthritis clinical trial: A case study. In *Communications in statistics—Theory and methods* (Vol. 38, # 18, pp. 3282–3296).

Ting, N. (2011). Phase 2 clinical development in treating chronic diseases. *Drug Information Journal, 45*(4), 431–442.

Wang, X., & Ting, N. (2012). A proof-of-concept clinical trial design combined with dose-ranging exploration. In *Biopharmaceutical statistics.* wileyonlinelibrary.com. doi:10.1002/pst.1525

Yuan, G., & Ting, N. (2014). First dose ranging clinical trial design—More doses? Or a wider range? In *Clinical trial biostatistics and biopharmaceutical applications.* Abingdon: Taylor & Francis (to be appeared in 2014).

Chapter 7
Risks of Inconclusiveness

7.1 Introduction

During the entire clinical development of a new medicinal product, there are lots of milestones and decision points. Many of these key decisions could have long-term impact and involve a large amount of resources and investment. As indicated in previous chapters, the Go/NoGo decision after a PoC study is one of such examples. A NoGo decision means that all of the investments and efforts in developing this candidate up to early Phase II have not paid off. The high hopes of the many scientists and other project team members will not be realized. On the other hand, a "Go" decision implies a significant amount of additional investments and resources, as well as more man power will be committed in further development of this candidate. Important activities such as purchasing large amount of raw materials for preparing Phase III clinical supplies, as well as long-term toxicology studies, will be initiated.

However, such a Go/NoGo decision is not an easy decision. Usually it takes time to make a clear decision after the PoC study final data read out. When this is the case, the time between PoC results and a final decision is considered as a period of "inconclusiveness."

In addition to such major decisions, other Phase II decisions can be difficult as well. For example, how to choose doses for the upcoming clinical trial design? Which observed doses can be thought of as "efficacious"? What type of dose-response relationship can be considered in moving forward to the next step of development? In this chapter, the inconclusiveness (Ting 2011) regarding PoC decision will be discussed in further detail. Moreover, a few ambiguous situations in dose ranging and recommendations of finding minimum effective dose are also covered in this chapter.

One popular confusion of evaluating a clinical project team's success is the use of the candidate product's success. In other words, some people tend to say a team is successful because the product is successful. Accordingly, if a product fails, the

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_7

impression is that the team failed. This is, in fact, a misunderstanding. In clinical development of new drugs, the criterion to evaluate the success of a team could have nothing to do with the success of the product. An efficient project team should allow the good product to reach patients as soon as possible and should stop developing a potentially unsuccessful drug as early as possible.

On this basis, the criterion to evaluate the success of a team should be "re-work." In clinical development, re-work does not only cause delay in decision making but can also be very costly. The patent life of a product is 20 years. The time to develop a product can run at least 7 or 8 years and, in many cases, this time could be much longer. A re-work of one Phase II clinical trial can easily erode the patent life by at least one year. The amount of potential revenue of a successful product could run up to billions of dollars or more each year. Hence the delay caused by re-work can be very expensive. Major re-work such as adjusting for underestimation of maximal tolerable dose (MTD) could be considered as a failure of the project team, regardless on whether the product being developed will eventually be successful or not. The real example introduced in Sect. 5.3, which took three studies to eventually identify the dose-response relationship, can also be interpreted as a failure of the team.

7.2 Go/NoGo Decision in a Two-Group PoC Study

One of the most important contributions a statistician makes to the clinical development team is statistical consulting. This includes many types of communications —oral communication, teleconference, email exchange, written memos, statistical analysis plans, study reports, publications, and other types. Among all these communications, one key message is regarding the probability of risks. In a proof of concept study, the study results should indicate a "Go" decision or a NoGo decision regarding the product under development. Hence before the study results read out, it is critical that the statistician communicates clearly with the team regarding the probabilities of making a wrong decision—either a wrong "Go" decision (making a Type I error) or a wrong NoGo decision (making a Type II error).

However, in many situations, such a Go/NoGo decision may not be straightforward. This chapter attempts to describe some of these situations and proposes some tools to help with these difficult communications. As indicated in Chap. 4, a positive PoC result triggers many important development activities. A long-term toxicity study is one of those activities. Drug formulation and drug supply are also dependent upon a positive PoC. If the concept is not proven, there is no need to stock up a huge pile of raw materials to make a large quantity of study drug or study biologic. Hence there is a practical impact of not being able to make a Go or a NoGo decision after the data read out.

When a clinical statistician calculates sample size for a new study, the four pre-specified quantities needed are alpha, beta, delta (δ, the treatment difference) and sigma (σ, the within-group standard deviation) of the primary endpoint. In most

of the disease areas, the variance of the primary endpoint is generally well approximated, and it tends to be stable. However, the most difficult quantity to come up with is δ. In order to simplify the discussion that follows, we assume delta to be such that a more positive value indicates a more favorable outcome. Let this delta be the difference between test treatment subtracted from placebo treatment, where a higher score is considered more favorable. Furthermore, we assume (without loss of generality) the variance is 1 (in case variance is not one, then let δ be the treatment difference divided by pooled standard deviation). Let alpha (α) be the pre-specified one-sided Type I error and beta (β) be the pre-specified Type II error (so that power is 1 − β).

Agreeing on a clinically meaningful treatment difference from placebo control is a very difficult problem in the early Phase II clinical trial design. This is one of the greatest challenges at this stage. However, without a delta, it would be virtually impossible to calculate the sample size for a clinical trial. Once the delta is obtained, the next natural step is to discuss with the team the different types of risks the team will take, namely, the selection of a Type I error rate and a Type II error rate.

Now, let z_α be the positive value of critical point on standard normal distribution with α on the right. For example, if α is 0.025, then $z_\alpha = 1.96$, and if α is 0.05, then $z_\alpha = 1.645$ (Fig. 7.1). Similarly, let z_β be the absolute value of critical point on standard normal distribution with β to the left of the alternative distribution (e.g., $z_\beta = 1.04$ when $\beta = 0.15$). Then, for any given alpha and beta, the sample size of each treatment group is $n = 2 * ((z_\alpha + z_\beta)/\delta)^2 = 2 * (Z/\delta)^2$ (denoting $Z = z_\alpha + z_\beta$). Note from Fig. 7.1 that z_α is associated with the null distribution on the left, and z_β is associated with the alternative distribution on the right. Let μ_P denote the mean of the placebo treatment and let μ_T be the mean of test product (higher values are more favorable). Then the hypothesis

$$H_0: \mu_T \leq \mu_P \text{ versus } H_1: \mu_T > \mu_P$$
$$(\text{which can be re-written as } H_0: \delta \leq 0 \text{ versus } H_1: \delta > 0)$$

is tested at level α.

A typical PoC clinical trial is designed with such a sample size, involving a comparison of a high dose of the product candidate with placebo to examine whether statistical significance of the treatment difference is greater than δ. In fact, μ_P (mean of the placebo group) is on the left, μ_T (mean of the test treatment group) is on the right, and the difference between μ_T and μ_P is δ. Then, by going through the standardization process—subtracting μ_P and dividing by the standard deviation (assumed to be 1 here for each of the two treatment groups), then the relationship looks like Fig. 7.1.

From Fig. 7.1, the distance between z_α and δ reflects the absolute value of z_β. Hence $\delta = z_\alpha + z_\beta$. After the study is completed, clinical data are analyzed and results are presented to the project team. The team members and upper management evaluate these results and make a decision This is the time when inconclusiveness could happen. A Go/NoGo decision is easy if both statistical significance is

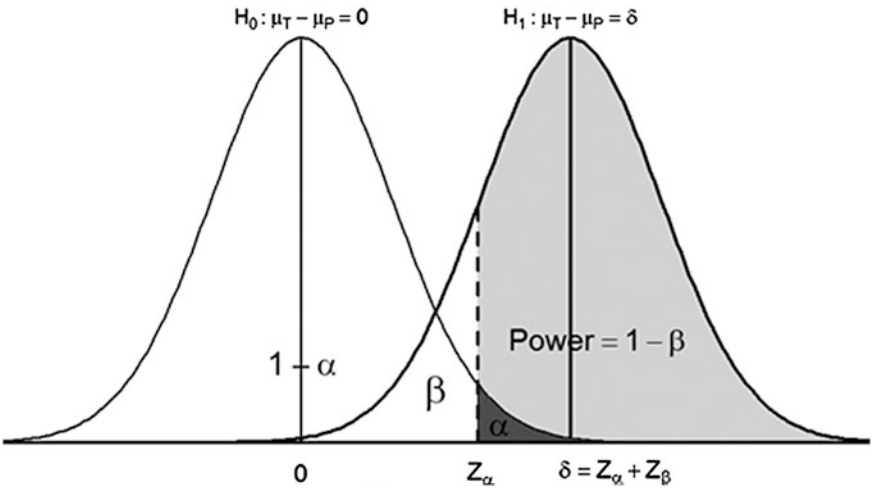

Fig. 7.1 Standard normal distributions under the null and alternative hypotheses

observed and treatment difference >δ is achieved. These results indicate a clear "Go" decision. On the other hand, when the results are not statistically significant, and the observed treatment difference is too small, then the team may make, often times reluctantly, a NoGo decision. However, if the observed treatment difference is not too small, but failed to reach statistical significance, then the team is in a difficult position and not clear as to whether a "Go" or a NoGo decision should be made. Under this circumstance, a situation of inconclusiveness may take place.

A typical early phase clinical trial tends to be designed with a relatively short duration. Many Phase I trials are single-dose trials, which means only one single dose is given to the trial participants. Some other trials may be designed with treatment duration of one week or two weeks. The limiting factor of time length comes from the pre-clinical toxicity studies.

In many cases, when the observed *p*-value is close to statistical significance, but still not statistically significant, the team tends to decide to move the candidate product forward. Also, in most of the cases where the observed $\widehat{\delta}$ value is between z_α and δ, the decision is to "Go." However, future development could indicate that this "Go" decision was a mistake because the observed $\widehat{\delta}$ was not large enough. These difficulties occurred because, at the time of designing PoC, a δ which is larger than the actual (true) treatment difference was used for various reasons. One typical case could be that such a δ was selected based on the available sample size; that is, from the sample size and the given α and β, a δ was calculated. A larger δ may have been chosen so that such a PoC study was feasible under the given budget constraints. However, after the PoC results are ready, although the observed $\widehat{\delta}$ was actually smaller than δ, the project team still made a "Go" decision simply because the *p*-value was either statistically significant or close to be significant.

On the other hand, under certain situations, when the observed $\widehat{\delta}$ is to the right of z_α (but to the left of δ), yet a NoGo decision was made. This is the case where a statistical significance was observed, but the clinically meaningful difference was not achieved. One example can be found when the PoC endpoint is different from the Phase III endpoint. When this is the case, a lowest δ was selected for sample size calculation during the study design stage. Given this background, the decision rule was that, if the study results cannot deliver such a δ, then there is no hope for the test product to meet the Phase III primary endpoint.

Under certain other situations, the PoC study may have to be repeated so that a clear decision can be made. Inconclusiveness can be thought of as the time period between the study results read out and a definitive Go or NoGo decision is made. In practice, when the study results are ready, the team decision does not necessarily follow the statistical criteria. There are cases where the team made a Go decision, even when the observed *p*-value is greater than α (this is the case that α is inflated). There are also cases that when the observed *p*-value is less than α, and the observed $\widehat{\delta}$ is less than δ, but a NoGo decision could have been made (β could be inflated). In either case, the risk of making a wrong decision is greater than the pre-specified risk.

Based on the above discussion, it is clear that when the observed $\widehat{\delta}$ is greater than z_α, and the decision is Go, then the risk of making a wrong decision is preserved under the pre-specified level. Also, when the observed $\widehat{\delta}$ is less than z_α, and the decision is NoGo, then the risk of making a wrong decision is also preserved under the pre-specified level. However, if a Go decision is made when the observed $\widehat{\delta}$ is less than z_α, then the risk of making a wrong Go decision is increased to be more than the pre-specified risk level. On the other hand, when the observed $\widehat{\delta}$ is greater than z_α, and a NoGo decision is made, then the risk of making a wrong NoGo decision is increased to be more than the pre-specified level β.

Of course, these decisions are usually made from long discussions among team members after looking at all clinical data on hand—primary and secondary endpoints, multiple doses if more than one dose was used, safety data, and sometimes pharmacokinetics (PK) data. The main issue to be addressed here is during the time between the PoC data is ready and the actual decision is made, a period of time that the PoC study was considered as "inconclusive". This time period is difficult because there is always tight timelines for the entire project. Will there be a need for a long-term toxicity study? If so, how long? And when should it start? What amount of drug supply will be necessary? At what time frame (Phase II requirements and Phase III requirements)? What dosages to be formulated? Will there be additional PK studies such as drug-drug-interaction or food effect studies to be designed?

At this time, a Go/NoGo decision is very critical. However, the data are not always clear. Regardless of all these discussions and examination into multiple angles or perspectives, by the end, there will have to be a decision—to go or not to go. This problem becomes much more difficult when the PoC study needs to be repeated. Hence we consider this period as the time of "inconclusiveness."

Table 7.1 For the observed $\widehat{\delta}$, and the team decision, and the associated consequences

Scenario	Observed $\widehat{\delta}$	Decision	Consequence
1	$\widehat{\delta} > z_\alpha + z_\beta$	Go	Study results meet both statistical significance and clinically meaningful difference between the two treatment; under this situation, the potential Type I error is much smaller than α
2	$z_\alpha < \widehat{\delta} < z_\alpha + z_\beta$	Go	Type I error is controlled under α, however, the clinically meaningful treatment difference is not achieved
3	$z_\alpha < \widehat{\delta} < z_\alpha + z_\beta$	NoGo	Type II error is inflated
4	$0 < \widehat{\delta} < z_\alpha$	NoGo	There is no inflation of Type II error
5	$0 < \widehat{\delta} < z_\alpha$	Go	The team inclined to make a "Go" decision, knowing that Type I error is inflated, this is the case where clear communications of risks are necessary

7.2.1 The Decision Process

After study results are ready, the decision would be very easy if either the observed $\widehat{\delta}$ is very large (an easy "Go" decision) or the observed $\widehat{\delta}$ is very small—very close to zero, or even negative (then a NoGo decision is made). However, this decision becomes more difficult if the $\widehat{\delta}$ is greater than zero but not sufficiently large. Table 7.1 listed some possible scenerios. In Table 7.1 for a few possible scenarios, with the observed $\widehat{\delta}$, given that the team makes such a decision, then the consequences are listed in Table 7.1.

In practice, situation #3 above could take place and the team is willing to accept a larger Type II error. However, situation #5 above becomes a challenge to the team and to the statistician. It is important to note that this communication should take place before the blind is broken because, after clinical results read out, the team will focus on the data, and it would be too late for the statistician to inform and guide the team regarding risks. Under the scenarios proposed, the statistician would be able to communicate the associated risks with the team before it is too late (after the blind is broken).

7.2.2 The Concept of Another Delta

The team may want to consider a different delta to aid in the decision-making process regarding the Go/NoGo decision. The statistician could use this concept as a communication tool to discuss the additional risks the team is taking, that is, how much of additional Type I error risk this decision may introduce.

If the Go/NoGo decision is simply based on the observed *p*-value, then the decision is not hard to make. In practice, there are many meetings and discussions after the study is un-blinded, with statistical results are reviewed by various stake holders. Typically, if the *p*-value is less than the pre-specified alpha, the secondary endpoints demonstrate a similar trend, and if there is no excessive safety concerns, then the team tends to make a Go decision so that this candidate will continue to be developed.

Nevertheless, when the *p*-value is slightly greater than alpha, only some secondary endpoints demonstrated good efficacy, and the safety profile is acceptable, should a Go decision be made? At this point, one important question would be "how large a *p*-value should be so that a definite NoGo decision will be made?" Or alternatively, the question could be "how small an observed $\widehat{\delta}$ should be so that a definite NoGo decision will be made?" Based on this line of thinking, the discussion between the statistician and the project team tends to be that "is there another quantity, denoted as δ', such that when the observed $\widehat{\delta}$ is less than δ', then the decision will be definitely NoGo. Therefore, after the sample size has been calculated based on the pre-specified delta, alpha and beta, a useful discussion would be to propose an alternate delta (δ') so that if the observed $\widehat{\delta}$ is less than δ', a clear NoGo decision will be made for this candidate.

Chuang-Stein et al. (2011) denoted the δ as the target value (TV) and the alternate δ' as lower reference value (LRV). They then propose a set of decision rules based on these two values. While the proposal has merit, there would also be a need to communicate with non-statisticians and use easy-to-understand decision rules. The statistician in the project team can simply lay out the strategy that if the observed $\widehat{\delta}$ is greater than δ (TV), then a Go decision will be made. If the observed $\widehat{\delta}$ is less then δ′ (LRV), then a NoGo decision will be made. Under this simple setting, it is obvious that when the observed treatment difference is between TV and LRV, then a situation of inconclusiveness may occur.

Note that the purpose for pointing out the risks of inconclusiveness is not to solve this problem but instead to educate the team that this is a true risk existing in the real-world drug development process. It is the statistician's role to help point this out to the project team, which will allow the team to propose strategies in order to handle their specific concerns. Complicated statistical procedures may not necessarily be practical in the application of individual cases. Statisticians are encouraged to employ a simple tool of communication so that non-statistical project team members, as well as upper management, can appreciate the existence of such risks.

7.3 Go/NoGo Decision with Multiple Treatment Groups

As discussed in Chap. 6, when a PoC study is combined with a dose-ranging clinical trial, one proposal is the use of trend test to establish PoC, and then to perform estimation for finding individual doses as compared with the placebo

control. In this proposal, the PoC step is accomplished with a single degree of freedom trend test. In a study design with a total of two groups or three groups (one test dose or two test doses plus placebo), the trend test is equivalent to a simple PoC. That is, test the null hypothesis that the high dose against placebo at α, if this null hypothesis is rejected, then claim the concept is proven. Otherwise claim the concept is not proven and a NoGo decision should be made. In these simple cases, the risks of inconclusiveness are the same as those discussed in the above section.

When more than two doses are included in a PoC combined with dose-ranging design, potential inconclusiveness could occur. In this case, the trend test serves as the single step for a PoC decision. If the null hypothesis of the trend test is accepted, and the monotonicity assumption remains true, then this result implies that none of the test doses delivers a treatment difference from placebo being greater than δ. The project team may opt for a NoGo decision. Under this circumstance, one potential consideration that leads to a Go decision could be that the MTD was underestimated.

On the other hand, the trend test could demonstrate a statistical significance, yet none of the observed dose delivers a meaningful treatment effect δ, as compared with the placebo control. Such a study result also creates risks of inconclusiveness. Under this situation, if a NoGo decision is made, then the Type II error (β) could be inflated based on the designed criteria. Nevertheless, a Go decision could be difficult because a clinically meaningful treatment difference is not observed from any of the test doses.

If the high dose is safe, this result could also be a potential case of underestimation of MTD. As indicated in Chap. 4, the MTD is obtained from Phase I results. If MTD is overestimated, then excessive toxicities could be observed in Phase II. However, if MTD is underestimated, then it may take additional studies to help revise the previous MTD estimates. This causes re-work, additional investment, and time delay in clinical development. But if the MTD is correctly estimated, and still no dose delivers a treatment difference of δ against placebo, then this becomes a case of inconclusiveness. One proposal of handling this situation could be to require at least one dose to deliver a statistically significant difference from placebo.

Suppose the MTD is correctly estimated, the dose-response relationship appears to be monotonic, and the null hypothesis that there is no dose trend is rejected. Under this circumstance, if none of the tested dose delivers a treatment difference greater than δ, the proposal is to employ a gatekeeping procedure (Wang and Ting 2012), in addition to the trend test. Basically this proposal is to design such a combined PoC and dose-ranging trial with a trend test plus a gatekeeping procedure. The first step is the trend test; if significant, then test the highest dose against placebo, with the entire alpha. If the test result is statistically significant, then conclude the concept is proven, and move down to test the next dose with the entire alpha. If the highest dose is not statistically significant, declare the concept is not proven. If this strategy is considered at the design stage, then the time of inconclusiveness could be minimized.

7.4 Dose Titration Studies Cannot Be Used for Dose-Finding

The dose-escalation studies implemented in Phase I are fundamentally different from dose titration studies implemented in later phases. The main difference is that in dose-escalation studies the different cohorts of patients are exposed to different doses of the study medication. On the other hand, in dose-titration studies, patients are randomized to treatment groups where the same patient is exposed to a set of different doses. For example, in a dose-titration design with test drug compared against placebo, if the test doses under consideration include 10 and 20 mg, then a titration design randomizes patients into two treatment groups—test drug against placebo. For patients randomized into the test drug group, they started with 10 mg. Over time, though, patients may be exposed to low dose and high dose, or for those patients randomized to placebo, they are exposed to the corresponding placebo treatment.

Dose titration designs include forced titration and response guided titration. For example, in an eight-week study with forced titration of test drug against placebo, with 10 and 20 mg dosages, the design may dictate that every patient starts with 10 mg of test drug or matching placebo. Then after two weeks of 10 mg (or placebo) double-blind treatment, every patient will be titrated up to 20 mg until end of treatment (week eight). On the other hand, a response-guided titration design (again, two treatment groups with test drug against placebo, and eight weeks duration of treatment) would randomize patients into one of the two treatment groups, and every patient starts with 10 mg (or matching placebo). Then, depending on patient responses, if the efficacy is not sufficient, the treating physician may consider titrating the dose up to 20 mg. However, after the patient on the 20 mg dose, if there is safety concerns, then the physician could down titrate the dose back to 10 mg. Hence dose-titration designs can be viewed as two different types: forced titration and response-guided titration.

In a forced titration study, this design may have been dictated by safety experience from previous studies. For example, if patients cannot tolerate a 20 mg dosing at the beginning, there could be many adverse events, and patients may be dropped out of study within the first 2 weeks of treatment. Then the design starts patients with 10 mg so that patients can develop tolerability of this test drug. Later after two weeks, if the patients' systems are tolerating the test drug, then the dose is titrated up to 20 mg in order to achieve the anticipated efficacy.

After the study completes, final statistical analysis is applied to test the hypothesis that whether the two treatments (test drug vs. placebo) are the same. Suppose the results are statistically significant, indicating that the observed efficacy of the test drug is different from that of the placebo. This result provides evidence to support that the designed regimen is effective, but there is no data to support dose-response relationship. In other words, the only conclusion is that the studied regimen (two weeks of 10 mg dosing followed by six weeks of 20 mg dosing) is

superior to placebo. But this result cannot be used to differentiate patients' response to 10 mg dosing from 20 mg dosing.

In the response-guided dose titration study, the dosing picture is more ambiguous. Suppose in a eight-week study with two treatments (test drug against placebo), the test drug can be titrated between 10 and 20 mg. If at end of study the statistical analysis indicating a significant treatment effect, then the project team concludes that there is a treatment difference. However, no inference about dose-response can be formulated from these results. If the project team attempts to tease out dosing information based on the last dose each patient is taking, then there could be interesting outcomes.

For example, in previous experiences, it could happen that such an analysis provides wrong dose-response relationship in both efficacy and safety. The subgroup of patients took 10 mg as the last dose demonstrate better efficacy than the subgroup taking 20 mg as the last dose. Also, the 20 mg subgroup of patients developed less adverse events than the 10 mg subgroup. This situation occurred because during the study those patients who responded favorably to the 10 mg dose would have stayed at the 10 mg dose. Only those patients who did not respond at 10 mg were titrated to 20 mg. Unfortunately, these patients still are not responding at 20 mg.

Hence, by end of study, the 20 mg subgroup of patients includes many of these non-responders, while most patients in the 10 mg subgroup are the responders. As such, it appears that the 10 mg subgroup delivered better efficacy than the 20 mg subgroup. Similarly, from a safety point of view, those patients who tolerate 20 mg dose get to stay at 20 mg. Only those patients who could not tolerate 20 mg dose were down titrated to 10 mg. In this case the 20 mg subgroup of patients look safer than the 10 mg subgroup Therefore, if the development objective is finding dose-response relationship, application of a dose-titration design could lead to another form of "inconclusiveness."

7.5 A Practical Design to Help Finding MinED

The challenges and difficulties relating to finding MinED (ICH E-4) are discussed in Chap. 5. The key lesson is that if MinED was not detected at Phase II, it will be too late and too expensive to find it in later phases. Minimum effective dose (MinED) is a very important concept in product development. Unfortunately, there is not a universally accepted definition for this term (Thomas and Ting 2008). Finding MinED is critical because it is generally believed that toxicity increases as dose increases. Hence a lower dose that is effective implies it could be safer than higher doses.

In proposing early Phase II dose-ranging clinical trial designs, the coverage of a wide range of doses is much more important than the consideration of dose-response models. In practice, selection of dose range starts from the high dose —either use of MTD or a dose that is not far below MTD. For the most common

cases, this dose tends to be the highest available dose under existing formulation or multiples of such a high dose formulation. The design process follows to select doses lower than this highest dose and gradually moves down to the lowest dose. In such a design, the dose range is then calculated as the highest dose divided by the lowest dose.

Thus, practically, selection of doses in a dose-ranging design is not from low dose moving up to high dose but rather from high dose moving down to low dose. When proposing a dose-ranging design with a wide dose range, the question is not "How high can you move up?" It is "How low can you go down?" On this basis, binary dose spacing (BDS) proposed in Chap. 5 is a very practical recommendation (Hamlett et al. 2002). Under the BDS framework, when more doses are included in a design, lower doses will be selected based on the available formulated dosages.

There are many advantages of selecting a wide dose range. In other words, a wide dose range indicates that the low dose in the given design can go very low. When this is the case, if the lowest observed dose delivers sufficient efficacy, then such a dose would be relatively safe because it is far lower than the MTD. On the other hand, if the lowest dose does not deliver efficacy, team members look for a higher dose that delivers efficacy. Once such a dose is found, then the observed lowest dose that provides meaningful efficacy could usually be accepted as MinED, because the observed data indicate that a dose lower than the perceived MinED is not efficacious.

For instance, in the motivating example presented in Sect. 5.3, it took three dose-ranging studies to identify the dose-response relationship. In that example, the first design includes placebo, 80, 120, and 160 mg of test drug. The second design includes placebo, 40, 80, and 120 mg doses. The third study uses placebo, 2.5, 10, and 40 mg. If there is an opportunity to re-do this development program using the current knowledge such as 4 or 5 test doses and BDS, then the first study would have been designed as placebo, 160, 40, 10, and 2.5 mg of test drug. Study results of such a design would indicate that 2.5 mg is not efficacious and that efficacy responses of 40 mg are very close to that from 160 mg. And, most importantly, clinical researchers can accept that 10 mg as the MinED with good practical justification.

Therefore, the popular misunderstanding of finding MinED is that researchers tend to struggle with a definition of such a dose, and statisticians are confused with selecting a "good algorithm" for finding MinED. In fact, the practical way of thinking about MinED should start from a wide dose range. In other words, for every dose ranging design, the key question really should be "How low can you go?" Practically, this question is limited by formulation. Therefore in preparations of early Phase II clinical trials, it is desirable to formulate some very low doses, so that a wide dose range can be explored.

7.6 Discussion

The research and development of new medicinal products is a very expensive and time consuming process; it is especially difficult during the clinical development stage. A Go/NoGo decision and dose-finding practices are extremely challenging. One important problem that deserves more attention would be the risks of inconclusiveness. In this chapter, a few situations of inconclusiveness are introduced, and some communication methods are recommended to help minimize the time wasted during the inconclusive stage.

In the case when a PoC is combined with dose-ranging trials, then the proposed approach from Chap. 6 would be to start with a trend test, and then followed with estimation of dose-response relationship. However, in case the trend test turned out to be statistically significant, but none of the observe dose delivers a meaningful difference from placebo. Then the inconclusiveness could be developed. When there is concern of this potential inconsistency, the recommendation is to add a gate-keeping procedure after the trend test. Hence the concept can only be considered as proven if both the trend test is statistically significant and the highest dose is also significantly different from placebo, in a sequential fashion (trend test has to be significant first, followed by the high dose being significantly different from placebo).

It is suggested that dose-titration designs are not used for dose-finding purposes. Regarding the difficulties in finding MinED, the proposal is to design the first dose-ranging study with a wide dose range, using 4 or 5 test doses. Binary dose spacing is a flexible and practical design tool. If the dose range is wide enough, there is a good chance to catch one or more low doses that appear to be sub-therapeutic, not efficacious. Then the lowest observed efficacious dose could be considered as a candidate of MinED.

Experiences suggest that whenever an important mistake is made in the clinical development of a new medicinal product, the critical point was not because "We did not know" but rather, in fact, "We thought we knew." Conducting clinical trials is taking observations from live human being, and hence there are lots of ethical concerns. Declaration of Helsinki (World Medical Association 2013) provides a set of ethical principles regarding human experimentation. In performing clinical trials, one of the most important scientific justifications is that "because we do not know whether the drug works or not, we need to study this in human body."

In other words, if we know the drug works, then it is not ethical to run clinical trials. In designing trials, people usually make assumptions. Some assumptions are more realistic, and some of the assumptions may not be realistic or practical. The misunderstanding at this stage is that some team members took the assumptions as known facts and moved forward. Hence it is important that when designing any clinical trials, the scientists remain humble and open-minded, and not make unwarranted assumptions.

As mentioned in Sect. 7.1, the metric to measure a clinical team's performance should not be relating to the success of the candidate under development. The more appropriate metric would be the amount of "re-work." If most of the team members keep this metric in their mind when designing or conducting clinical trials, it is hoped that the inconclusiveness can be reduced to an acceptable amount. Again, the inconclusiveness of Go/NoGo decision could potentially cost the sponsor lots of investment and waste precious development time. It is the team members responsibility to keep it to a minimum.

References

Chuang-Stein, C., Kirby, S., French, J., Kowalski, K., Marshall, S., Smith, M. K., et al. (2011). A quantitative approach for making Go/NoGo decisions in drug development. *Drug Information Journal, 45*(2), 187–202.

Hamlett, A., Ting, N., Hanumara, C., & Finman, J. S. (2002). Dose Spacing in early dose response clinical trial designs. *Drug Information Journal, 36*(4), 855–864.

Thomas, N., & Ting, N. (2008). Minimum effective dose. In R.B. D'Agostino, L. Sullivan, & J. Massaro (Eds.), *Encyclopedia of clinical trials*. Hoboken, NJ: Wiley-Blackwell.

Ting, N. (2011). Phase 2 clinical development in treating chronic diseases. *Drug Information Journal, 45*(4), 431–442.

Wang, X., & Ting, N. (2012). A proof-of-concept clinical trial design combined with dose-ranging exploration. *Pharmaceutical Statistics*. wileyonlinelibrary.com. doi:10.1002/pst.1525

World Medical Association. (2013). Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA, 310*(20), 2191–2194.

Chapter 8
Analysis of a Proof of Concept Study

8.1 Introduction

From the data analysis perspective, clinical statisticians are typically faced with continuous data, discrete data, or time-to-event data. For early Phase II clinical trials, the most frequently seen data types are continuous data or binary data. In this chapter, and the next two chapters covering data analysis, the discussion will mainly be about analyzing continuous data or binary data.

In typical clinical trials, continuous data tend to follow normal distribution, log normal distribution, or neither. When the continuous data are neither normally distributed nor log normally distributed, nonparametric analysis can be applied to handle these situations. One of the popular nonparametric approaches could be to rank these data first and then to perform parametric analysis on the ranked data. The central limit theorem for means, one of the most celebrated results in statistics, makes the assumption of normality appropriate on the sampling distribution of the mean even if the individual data are not normally distributed, provided that the sample size is large enough (say, 30 subjects per treatment group). Therefore parametric statistical tests for means, which assume normality, are appropriate under these circumstances. When the sample size is not large enough, nonparametric alternatives should be considered. Sample sizes of many Phase II clinical trials exceed 30 subjects per group.

Give this background, analysis of continuous data obtained from Phase II clinical trials are typically performed using analysis of variance (ANOVA), analysis of covariance (ANCOVA), or mixed effects model with repeated measures (MMRM). If the continuous data follow log normal distribution, then the raw data are first log-transformed, and statistical analyses are applied on these log-transformed data. Similarly, if the continuous data are neither normally distributed nor log normally distributed, then rank transformation can be applied to the raw data and then ANOVA or ANCOVA can be used to analyze these rank-transformed data.

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_8

In medical practice of drug treatment, a patient is usually classified as a "responder to the treatment," or a "non-responder" based on follow up observations obtained while or after this patient is treated. Hence clinical trial outcomes from each patient can be summarized into a responder status—a patient either "responded" or "did not respond." Hence binary data is another popular analytical data type (in addition to continuous data). In the analysis of binary data from a proof of concept study, the outcomes can be summarized into a 2×2 table: two treatment groups (test drug vs. placebo control) with two possible outcomes (responder status: Yes or No). Typical statistical analytic tools for binary data involve a chi-square test or logistic regression.

In this chapter, statistical analyses of PoC study results are presented and discussed using examples. Based on the traditional PoC study design, only two treatment groups: a high dose of test-treatment group and a placebo-control group—are included in the PoC study. The two types of primary endpoints considered are in the form of either continuous data or binary data. Details of their analyses are given below, with two examples provided.

8.2 When the Primary Endpoint Is a Continuous Variable

8.2.1 Data Description and Hypothesis

In the following illustrative (hypothetical) example, a Phase II proof-of-concept trial is designed to assess a test agent in reducing a type of chronic pain as compared with a placebo control. The primary measure is a patient-rated 0–100 visual analog pain scale, a measurement instrument intended to measure pain across a continuum from none (0) to the most extreme amount of pain (100). Here lower values indicate less pain (more favorable). Treatment period is eight weeks with weekly visit for the first 4 weeks, plus a week 6 visit and a week 8 visit—a total of six visits denoted as visit 1, 2, 3, 4 (corresponding to the first 4 weeks), visit 5 is the week 6 visit, and visit 6 as the week 8 visit. Baseline pain is measured at visit 0,

Table 8.1 Summary statistics for pain scale at each visit

Time	Placebo			Test drug		
	Sample size	Mean	Standard deviation	Sample size	Mean	Standard deviation
1	95	23.6	6.0	98	22.7	6.5
2	91	20.0	6.6	92	20.4	6.7
3	86	19.8	6.4	92	18.9	6.9
4	81	18.4	6.0	89	17.4	6.7
5	75	16.5	6.6	85	14.6	6.0
6	71	14.1	6.5	84	12.2	6.9

before each patient is randomized. The pain scale is measured at every visit by each participated patient. Summary statistics are presented in Table 8.1.

This study involved 95 subjects randomized into the placebo-treatment group, and 98 subjects randomized into the test-treatment group. At baseline, the 95 placebo-treated subjects experienced a mean pain score of 24.4, with a standard deviation of 6.0; 98 test-treated subjects experienced a mean pain score of 25.6, with a standard deviation of 7.4 (Table 8.1). After 8 weeks of double-blind treatment, 24 placebo-treated subjects dropped out and 71 subjects provided assessments on pain using the 0–100 pain scale. For the 98 subjects randomized into the test treatment, 14 subjects dropped out and 84 subjects provided week 8 measures.

The primary statistical hypothesis for this clinical trial is

$$H_0 : \mu_{T8} = \mu_{P8} \text{ versus } H_0 : \mu_{T8} < \mu_{P8}$$

The above hypothesis is to be tested at the pre-specified one-sided alpha of 0.025. Here μ_{T8} stands for the mean change in pain from baseline to week 8 for the subjects treated with the test drug, and μ_{P8} stands for the mean change in pain from baseline to Week 8 in pain for those placebo-treated subjects up to week 8.

8.2.2 T-*Test Approach*

Under the intention-to-treat (ITT) principle, a clinical trial analysis should in principle include measurements on all patients and the clinical data should be analyzed as randomized. In the case of patient drop out, one common analytical methods is to carry forward the last available observation (last observation carried forward, LOCF) and perform a statistical analysis on these LOCF data under the assumption that the patient's score remains stable (Ting 2000). Even this method is somewhat debatable, one advantage of the LOCF analysis is that this concept is easy for non-statisticians to understand. LOCF also provides each patient an equal weight (one patient, one vote). The simplest statistical analysis on these LOCF data would be a two-sample t-test. The SAS code (version 9.4) of the *t*-test for differences from baseline (i.e., DIF) to LOCF are obtained as follows:

```
PROC TTEST DATA=DTA22;
CLASS TRT;
VAR DIF;
RUN ;
```

The analytical results from the above SAS program can be outputted as follows:

trt	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	95	−7.8734	7.3661	0.7557	−21.2584	14.0913
2	98	−11.7935	6.7892	0.6858	−26.0588	10.4313
Diff (1–2)		3.9201	7.0790	1.0192		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	191	3.85	0.0002
Satterthwaite	Unequal	188.61	3.84	0.0002

The first part of the output is the summary statistics to report the mean, standard deviation, standard error, minimum and maximum for each treatment along with the treatment means. The second part of the output is to report the results from the independent two-sample statistical *t*-test. The "pooled" is the *t*-test assuming equal variances and the "Satterthwaite" assuming non-equal variances (SAS Manual). The variances between these two treatment groups are similar and hence it is natural to consider the equal variance analysis from this two-sample *t*-test. The results from the above output show a *t* value of 3.85, with a significant *p*-value of 0.0002, which is much less than the pre-specified two-sided alpha of 0.05. Therefore, the null hypothesis is rejected and it can be claimed that the test drug is statistically significantly different from the placebo control in treating patients with this type of chronic pain.

8.2.3 Analysis of Covariance Approach

It is also very common to introduce covariates in data analysis. When this is the case, an analysis of covariance (ANCOVA) would be an appropriate model (ICH E9). The two widely accepted covariates to be used in ANCOVA are stratification factors (like center) and baseline measurements. In this hypothetical example, no stratification factor is used and hence the corresponding ANCOVA model only includes treatment effect and the baseline (a continuous variable) effect. In this analysis, the response variable is the observed difference in the visual analog pain scale from baseline to primary timepoint of the study (using LOCF for missing data). The SAS code for this ANCOVA is given as follows:

```
PROC GLM DATA=DTA22;
CLASS TRT;
MODEL DIF=TRT V0;
RUN;
```

The output for this analysis can be summarized as follows:

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Model	2	1964.30423	982.15211	22.35	<0.0001
Error	190	8348.39018	43.93890		
Corrected Total	192	10312.69441			

R-Square	Coeff Var	Root MSE	Dif Mean
0.190474	−67.20089	6.628642	−9.863920

Source	DF	Type III SS	Mean Square	F Value	Pr > *F*
treatment	1	581.168051	581.168051	13.23	0.0004
baseline	1	1223.030497	1223.030497	27.83	<0.0001

The first part of the above SAS output is the so-called ANOVA table to test the proposed overall ANCOVA incorporating baseline which is to report the sum of squares (denoted by `Sum of Squares`) and the mean square of errors (denoted by `Mean Square`) for both the ANCOVA model (denoted by `Model`) and the residual errors (denoted by `Error`), the *F* statistic (denoted by *F* `Value`) with the associated p-value (denoted by `Pr > F`). The *p*-value associated with the proposed ANCOVA is less than 0.0001 which indicates overall statistical significance. Then this model can be used to test the treatment significance which is reported as the last part of the output. Under the statistical hypothesis of interest, the observed *p*-value of 0.0004 indicates that the null hypothesis of no difference in means is rejected under a two-sided alpha of 0.05. This conclusion is consistent with that of the *t*-test.

8.2.4 Mixed-Effect Models to Analyze the Longitudinal Data

The same data set can also be analyzed using a longitudinal data model. Under the assumption of missing at random (MAR), one common model is the mixed-effects model with repeated measures (MMRM). In the illustrative MMRM (Mallinckrodt et al. 2008), the clinical endpoint collected at each time point is used as the response variable. Fixed effects include treatment (discrete) effect, time (taken as discrete) effect, baseline (continuous) effect, treatment-by-time interaction, and baseline-by-time interaction. Random effects include patient effects and the residuals. In this data set, V0 stands for the value observed at the baseline visit, and V6 is the observation taken at visit 6. The SAS code is given below:

```
PROC MIXED DATA=DTA21;
CLASS PID TRT TIME;
MODEL Y= TRT TIME TRT*TIME V0 V0*TIME / S ;
REPEATED TIME / SUBJECT=PID TYPE=UN R ;
ESTIMATE "V6" TRT -1 1 TRT*TIME =0 0 0 0 0 -1 0 0 0 0 0 1 ;
RUN;
```

The associated results from the above SAS code can be shown as follows:

Effect		Den DF	*F* Value	Pr > *F*
TRT	1	190	8.38	0.0042
TIME	5	190	9.70	<0.0001
TRT*TIME	5	190	1.96	0.0862
V0	1	190	160.86	<0.0001
V0*TIME	5	190	1.02	0.4049

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > |t|
V6	−2.6057	0.9148	190	-2.85	0.0049

It can be seen from the output that the observed *p*-value is 0.0049. Results obtained from MMRM are consistent with the *t*-test and the ANCOVA results in this particular case.

This hypothetical example is introduced to demonstrate how to analyze a set of continuous data in a two-treatment group PoC clinical trial. Continuous data (or at least data taken as continuous) are most commonly seen in PoC studies. Often times an arithmetic mean change from baseline to LOCF is considered as the primary endpoint, and the difference in mean changes between treatment groups is compared to test for statistical significance. In certain continuous primary endpoints (for example, lab data or pharmacokinetics data), the underlying distribution are considered as log-normal, and the interest is to analyze the ratio between treatments. When this is the case, the raw data are first transformed with log (or natural log) and then the t-test, ANCOVA, or MMRM is applied to the log-transformed data to perform the statistical analysis. After all of the analyses are performed, the resulted point estimates and interval estimates are then back transformed using the anti-log (or exponent) function. These back transformed analytical results are then reported for the clinical team to review and interpret.

In some other PoC studies, the continuous data do not follow a clear normal or log-normal distribution—the distribution may look skewed, there may be some outliers, or for various other reasons. When this is the case, analysis of the ranked data can be helpful when the sample size per group is small (say, less than 30). One proposal is to take the standardized rank of the original data and then perform statistical analyses on these standardized ranks. Standardized ranks are obtained by first ranking all of the data across treatment groups. These ranks are then divided by the total number of observations, resulting in ranks between 0 and 1. Standardized

ranks can be obtained from SAS, PROC RANK, using the FRACTION option. Under this circumstance, the *t*-test or ANCOVA with LOCF can be appropriate analytical tools when data are missing completely at random..

8.3 When the Primary Endpoint Is a Binary Variable

8.3.1 Data Description and Hypothesis

In this section, another hypothetical example is presented and discussed. This time the example is based on a PoC clinical trial where the primary endpoint is the number and percent of responders in each treatment group (Table 8.2). Again, there are two treatment groups: the test treatment group and the placebo control. Randomization is stratified by sex (male or female). This is a four-week trial and efficacy measures are collected at baseline and weekly at every week post-baseline. The primary endpoint is to compare the last observation of each patient against the associated baseline. If the change in clinical response achieves a pre-defined threshold, then this patient is considered as a responder. The patient would be denoted as a non-responder if his or her clinical measure fails to reach the pre-defined threshold.

In this hypothetical study, fifty-three patients are randomized to the test treatment group, and 56 patients are randomized to the control group. Study results are presented below. Cochran-Mantel-Haenszel (CMH) test is used to analyze the binary primary endpoint, using sex as a covariate. SAS output of the analytical results are presented below. The statistical hypothesis of interest is

$$H_0 : \pi_T = \pi_P \quad \text{versus} \quad H_0 : \pi_T \neq \pi_P$$

where π_T represents the proportion of responders in the test-treatment group and π_P represents the proportion of responders in the placebo-control group.

8.3.2 Cochran-Mantel-Haenszel Method

In SAS programming, it is important to note that under PROC FREQ, in the specification of TABLES statement, the order of variables is very critical. The FREQ procedure makes comparisons between rows and responses should be

Table 8.2 Primary efficacy binary data in hypothetical example

	Non-responder	Responder	Total
Placebo	45	11	56
Test drug	25	28	53

specified as columns. Hence response should always be the last variable to be given in the TABLES statement. Right before the response variable should be the treatment variable (the key comparison). All covariates need to be placed before the treatment variable. In this example, the SAS code can be as follows:

```
PROC FREQ DATA=DATA1 ;
TABLES SEX * TREATMENT * RESPONSE / CMH ;
RUN ;
```

The output from the above SAS FREQ Procedure is as follows:

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	5.4996	0.0190
2	Row Mean Scores Differ	1	5.4996	0.0190
3	General Association	1	5.4996	0.0190

Breslow-Day Test for Homogeneity of the Odds Ratios	
Chi-Square	0.7279
DF	1
Pr > ChiSq	0.3936

The observed *p*-value is 0.0190 (for pairwise comparisons, the *p*-value should be read from "Row Mean Score Differ"), indicating that there is a statistical significance in comparing the two proportion of responders. The Breslow-Day test of common odds ratio across strata presents a p-value of 0.39. Thus there is no statistical difference in responses between male patients and female patients, which can be thought of as lack of statistical interaction between treatment and sex.

8.3.3 Logistic Regression

This same data set can also be analyzed using the logistic regression. In the following set of SAS code and results, PROC LOGISTIC is used as follows:

```
PROC LOGISTIC DATA=DTA34 ;
CLASS TRT SEX ;
MODEL RESPONSE=TRT SEX ;
RUN ;
```

The logistic regression is fitted with maximum likelihood estimation and the likelihood ratio chi-square test is used to test for all effects from the regression effect which can be illustrated as follows:

Effect	DF	Wald Chi-Square	Pr > ChiSq
trt	1	5.4941	0.0191
Sex	1	1.6953	0.1929

From above results, the main comparison is based on the treatment effect (trt) and its p-value of 0.0191 indicates that there is a significant difference between the two treatment groups. Note that the CMH test provided a p-value of 0.0191. These two results are very close to each other.

8.4 Discussion

Although the examples in this chapter presented a variety of ways in analyzing the primary efficacy data from a PoC clinical trial, in reality, only one primary analytical method should be pre-specified from the protocol. For example, in the case of the illustrative pain study, if the primary analysis is pre-specified as the ANCOVA applied on LOCF, then the MMRM or *t*-test analysis can be considered as only exploratory. In order to preserve alpha, only one primary model can be used for the primary analysis using the primary endpoint. Allowing model selection will inflate alpha. Furthermore, this primary analysis applied to the ITT analysis set, at the primary analytical time point. Any deviation from the pre-specified endpoint, time point, analysis set, or analytical model could inflate alpha.

In some situations, one or more secondary endpoint(s) may also be considered in a PoC study. When this is the case, strategies of alpha allocation need to be clearly specified in the statistical analysis plan (SAP). The team should also agree, before blind is broken, as to how these secondary results will be used in the Go/NoGo decision.

In practice, the design of a PoC clinical trial is a difficult and challenging task (Chuang-Stein et al. 2011). After the data are ready, the primary statistical analysis of PoC results is relatively straightforward. However, a Go/NoGo decision may not necessarily be easy. In the previous hypothetical examples, statistical significance is achieved in both the pain study, and the study with binary primary endpoint. In fact, statistical significance may or may not be translated into clinical relevance. In many cases, even though a statistical significance is observed, the sample mean difference between treatment groups could still be less than the expected treatment difference that was used to calculate sample size. Under this situation, lots of discussions among team members and upper management may take place before a Go/NoGo decision can be made.

In addition to the primary endpoint, the team (and the upper management) need also look at secondary endpoints and safety profiles to help with the final decision. As indicated in Chap. 4, a “Go” decision after PoC triggers much development activities and substantial investment commitments. Sponsors do not take this decision lightly. On the other hand, a compound could have passed all the stringent

evaluations of the non-clinical development, formulation, toxicology, and all other hurdles before moving into clinical development. Furthermore, this compound will also have to endure the Phase I challenges before it can achieve PoC. This means that many scientists have high hopes about such a compound. A NoGo decision would, needlessly to say, be a disappointment.

It is because of all these situations, and other difficulties in decision-making as discussed in Chaps. 6 and 7, a PoC clinical trial has to be designed carefully, with much thought, consideration, and discussion. Nevertheless, after the primary analytical methods are well pre-specified, the statistical analysis of a PoC study is simple and straightforward. The primary endpoint of most PoC studies is either a continuous outcome or a binary outcome. Both cases are covered in this chapter with relevant examples.

References

Chuang-Stein, C., Kirby, S., French, J., Kowalski, K., Marshall, S., Smith, M. K., et al. (2011). A Quantitative approach for making Go/NoGo decisions in drug development. *Drug Information Journal, 45*(2), 187–202.

Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y., & Mancuso, J. P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal, 42,* 303–319.

SAS User's Manual, SAS Inc—SAS™ System (SAS Institute Inc., Cary, North Carolina), version 9.4.

Ting, N. (2000). *Carry-forward analysis, encyclopedia of biopharmaceutical statistics* (pp. 103–109). New York: Marcel Dekker, Inc.

Chapter 9
Data Analysis for Dose-Ranging Trials with Continuous Outcome

9.1 Introduction

Traditionally, proof of concept (PoC) study and dose-ranging trials are conducted separately in a sequential fashion. A two-group PoC design involves a comparison between MTD of test drug and a placebo control, a comparison that is conducted first. Only after the concept is proven, then dose-ranging trials are designed to study the dose-response relationship. In recent years, more sponsors are combining these two steps into one—using one trial with multiple doses to serve both the PoC and dose-ranging purposes. The focus of this chapter is to discuss data analysis of dose-ranging trials—either this dose-ranging trial is designed after PoC or in combination with the PoC objective. In data analysis it is very important to keep in mind that "analysis needs to respect design." Hence it is always a good practice to follow the pre-specified analysis, either from the protocol or from the statistical analysis plan (SAP).

In a combined PoC and dose-ranging clinical trial, the first step should be PoC based on statistical hypothesis testing. Chapter 6 recommends the use of a trend test to help addressing the PoC question. As discussed previously, if the study is designed with only two groups with high dose of test drug against a placebo control, then it is not a dose-ranging trial. For a clinical trial to serve as a dose-ranging study, it has to include multiple doses of the product candidate, plus control(s). In a multiple dose, dose-ranging study design, the first consideration is to see if PoC for this study drug has been established in a previous study. If so, then the current clinical trial is mainly designed to study the dose-response relationship. If not, then a PoC step would necessarily be the first statistical hypothesis to be tested. Chapter 8 covers data analysis under the two-group designs. This current chapter discusses data analysis when multiple dose groups are included in a single study.

In the analysis of dose-ranging clinical trials, there are typically three approaches commonly used to help determining the dose-response relationship: the first is a

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_9

multiple comparison procedure (MCP) approach, the second to use a dose-response modeling (Mod) approach and the third to combine both the MCP and Mod as a newer approach, which is commonly referred as "MCP-Mod" approach (Bretz et al. 2005; Pinheiro et al. 2006).

In the MCP approach, the dose is treated as a qualitative factor with a set of dose levels from the dosages available for testing in the clinical trial. In this approach, there are usually no assumptions about the underlying dose-response relationship and the statistical inferences to find doses of interest are restricted to this set of doses. In contrast to the MCP approach, the "Mod" approach typically assumes a parametric dose-response relationship between the dose and the efficacy response in the trial. Then the doses are treated as a quantitative continuous variable that would allow more flexibility for dose finding. Therefore, the validity of this "Mod" approach would depend strongly on the assumed dose-response model.

Combining both the MCP and Mod approaches, Bretz et al. (2005) developed an MCP-Mod approach to unify the MCP and Mod approaches, which is to assume the existence of several candidate parametric dose-response models and use multiple comparison techniques to choose the one most likely to represent the true underlying dose-response curve. At the same time, the hybrid MCP-Mod approach preserves the family-wise error rate. The selected model is then used to provide inference on adequate doses. Based on this novel approach, Bretz and colleagues developed an R package (i.e. "MCPMod") which is publicly available at http://cran.r-project.org/web/packages/MCPMod/index.html with detailed description in Bornkamp et al. (2009). Because thispackage is not further developed anymore. it has become part of the new package in "DoseFinding" which can be freely accessed at https://cran.r-project.org/web/packages/DoseFinding/index.html.

In this chapter, we illustrate how to use these R packages to analyze dose-response trials in a step-by-step fashion with detailed explanations so that the interested readers can use the packages and R programs in this chapter for their own analysis. In order to help readers better understanding data analysis of dose-ranging trials, we make use of the Phase II dose-finding study used in Bretz et al. (2005) as the example in this chapter, which was in fact part of the R packages with name "*biom*." We will use this data to illustrate the MCP, the Mod, and the MCP-Mod approaches.

This chapter is organized as following sections. Section 9.2 provides the data set and the associated descriptive statistics. Section 9.3 presents analysis for making a PoC decision. In the case when a previous PoC has concluded the efficacy of thestudy drug, readers may skip this section. The multiple comparison procedures (MCP) are introduced in Sect. 9.4 and the modeling (Mod) approach in Sect. 9.5. In Sect. 9.6, we introduce the combined MCP and Mod approach (MCP-Mod) and conclude the chapter with a discussion in Sect. 9.7.

9.2 Data and Preliminary Analysis

Let's first describe the data. As reported in Bretz et al. (2005), this study was a randomized double-blind parallel group trial with a total of 100 patients being allocated to either placebo or one of four active doses coded respectively as 0.05, 0.20, 0.60, and 1, with $n = 20$ per group. The response variable was assumed to be normally distributed with larger values for a better outcome. The data is public available in R packages "MCPMod" and "DoseFinding" as a data frame named "*biom*" with 2 columns named "dose" for dose levels and "resp" for responses. There are 20 observations for each of the 5 dosages. The data are reproduced here in Table 9.1. Note that the actual response data have 8 digits and are rounded to 3 digits in this table.

To analyze this data, we first load the data frame "*biom*" with the R library "MCPMod" or "DoseFinding." When the data are available for analysis, we can then call R function "*boxplot*" to make a boxplot for the data distribution as seen from the following R code chunk:

```
# Load the library
>library(MCPMod)
# Get the data
>data(biom)
# Get the mean response for each dosage with ''aggregate''
>aggregate(biom$resp, list(dose=biom$dose), mean)
# Call R function ''boxplot'' function to make the plot
>boxplot(resp~dose, biom, xlab=''Dose'', ylab=''Response'')
```

As seen from Fig. 9.1, there is an observed monotonic increasing dose-response relationship with sample means of 0.345, 0.457, 0.810, 0.934 and 0.949, respectively.

Table 9.1 Dose-response data from "biom"

dose	Resp
0.00	0.355 0.137 −0.017 0.367 0.395 0.358 0.313 −0.045 0.623 0.04 −0.308 −0.453 −0.202 1.561 0.561 0.336 −0.159 0.993 0.807 1.237
0.05	1.354 0.155 −0.071 0.584 0.963 0.381 −0.005 0.519 −0.363 0.339 0.146 0.443 −0.022 0.373 1.529 0.149 0.374 1.012 0.281 0.995
0.20	1.198 1.568 −0.156 0.207 1.853 0.996 2.452 −0.520 0.049 0.633 0.528 0.423 1.233 1.868 1.058 0.345 0.405 0.484 0.571 1.011
0.60	1.172 0.668 1.778 0.310 0.056 0.900 0.738 0.228 1.385 0.912 0.299 0.279 0.270 0.570 1.604 1.577 2.925 0.172 2.161 0.686
1.00	2.248 2.174 1.255 1.865 −1.113 1.198 1.975 0.625 1.330 0.149 0.887 0.563 0.731 −0.771 0.495 0.303 0.425 2.558 1.129 0.948

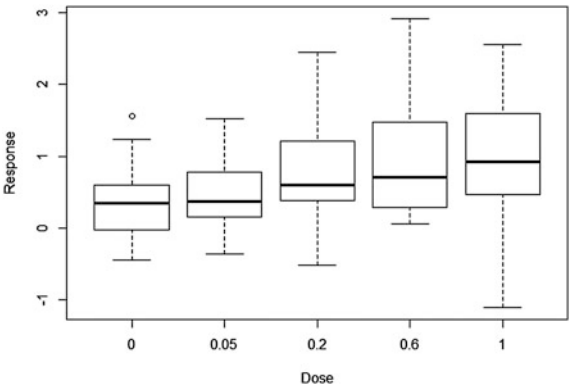

Fig. 9.1 Boxplot distribution for dataframe "biom"

9.3 Establishing PoC with a Trend Test

The PoC can be established based on the contrast test for statistical significance. Let μ_i be the population mean for dose group i (i = 0, 1, …, k), where μ_0 is the mean from the placebo group. For example, k is 4 corresponding to the "*biom*" data in Sect. 9.2 and the estimated sample means of μ_i (i = 0,1, 2,3,4) are 0.345, 0.457, 0.810, 0.934 and 0.949, respectively, as noted in the last section,

To construct a contrast for PoC, let c_i be the corresponding contrast coefficients with the condition that $\sum_{i=0}^{k} c_i = 0$. Then the null hypothesis (H_0) of no treatment effect versus the one-sided alternative hypothesis (H_A) of significant treatment effect (PoC) can be written as follows:

$$H_0 : L(\mu) = \sum_{i=0}^{k} c_i \mu_i \leq 0 \quad v.s. \quad H_A : L(\mu) = \sum_{i=0}^{k} c_i \mu_i = \delta > 0 \tag{9.1}$$

Corresponding to "biom" data with $k = 4$, the contrast coefficients are $c = (c_0, c_1, c_2, c_3, c_4) = (-2, -1, 0, 1, 2)$, which can be found in Table 9.2 in Wang and Ting (2012), and the above hypotheses become as follows:

Table 9.2 Commonly used dose-response models

Model	Functional form	Parameters
Linear	$E_0 + \delta d$	E_0, δ
Linear in log-dose (linlog)	$E_0 + \delta \log(d)$	E_0, δ
Exponential	$E_0 \exp(d/\delta)$	E_0, δ
Quadratic	$E_0 + \beta_1 d + \beta_2 d^2$	E_0, β_1, β_2
Logistic	$E_0 + \frac{E_{max}}{1 + exp((ED_{50} - d)/\delta)}$	E_0, E_{max}, δ
Emax	$E_0 + \frac{E_{max} d}{ED_{50} + d}$	E_0, E_{max}, ED_{50}
Sigmoid Emax	$E_0 + \frac{E_{max} d^{\lambda}}{ED_{50}^{\lambda} + d^{\lambda}}$	$E_0, E_{max}, ED_{50}, \lambda$

$$H_0 : L(\mu) = (-2)\mu_0 + (-1)\mu_1 + 0\mu_2 + 1\mu_3 + 2\mu_4 \leq 0, vs$$
$$H_A : L(\mu) = (-2)\mu_0 + (-1)\mu_1 + 0\mu_2 + 1\mu_3 + 2\mu_4 = \delta > 0$$

The test statistic can then be constructed as a one-sided t-test as follows:

$$T = \frac{\widehat{L(\mu)}}{SE\left(\widehat{L(\mu)}\right)} = \frac{\sum_{i=0}^{k} c_i \bar{x}_i}{SE\left(\sum_{i=0}^{k} c_i \bar{x}_i\right)} \tag{9.2}$$

where $\bar{x}_i$ is the sample mean of μ_i and SE denotes its standard error.

The step-by-step implementation in R can be done as follows. We first fit an analysis of variance (ANOVA) model to estimate the sample means and their associated standard errors:

```
# fit an ANOVA model to estimate means
>fit.aovmean = lm(resp~as.factor(dose)-1, data=biom)
# print the summary of fit
>summary(fit.aovmean)
```

In this fitting, we used R function "`lm`" (i.e., linear model) to fit the ANOVA model and then named it as "`fit.aovmean`." Notice that in this model fitting, we used an R function "`as.factor`" to change the numerical "dose" variable into a "factor" variable as categorical variable. Note that the "-1" in "`as.factor(-dose)-1`" is to tell the R function "lm" to fit a linear model without the intercept so to produce the estimated least-square means. Then we call the R function "`summary`" to print the summary of model fit, which can be outputted as follows:

```
Coefficients:
           Estimate Std. Error t value  Pr(>|t|)
(dose)0     0.3449     0.1593  2.165    0.0329
(dose)0.05  0.4568     0.1593  2.867    0.0051
(dose)0.2   0.8103     0.1593  5.087  1.83e-06
(dose)0.6   0.9344     0.1593  5.866  6.47e-08
(dose)1     0.9487     0.1593  5.956  4.35e-08


Residual standard error: 0.7124 on 95 degrees of freedom
Multiple R-squared: 0.5336,  Adjusted  R-squared: 0.509
F-statistic: 21.74 on 5 and 95 DF, p-value: 1.897e-14
```

We can see that the ANOVA model is statistically significant with p-value = `1.897e-14` and a R^2 = `53.36%`. The estimated means are labeled under the column "`Estimate`" for all the doses which are all statistically significant as shown for their associated p-values. The estimated standard deviation, also known as the residual standard error, is 0.7124. We can extract these values for PoC calculations as follows:

```
# Extract the estimated sample means
>EstMean.aov = coef(fit.aovmean)
# Extract the estimated sigma
>aov.sigma  = summary(fit.aovmean)$sigma
```

The statistical test for PoC in Eq. (9.2) can then be implemented in the following R code chunk:

```
# PoC: contrast mean >0 statistically?
>Dosage = sort(unique(biom$dose)) # get the dosages
>len.Dosage = length(Dosage) # the number of dosages
>alpha=0.05 # the alpha level
>n=20 # sample size for each dosage
# The c's for 4-dose and 1-placebo
>Contrast.Coef = c(-2,-1,0,1,2)
# Implementation of equation 9.2. with 1-sided test
# where ''qt'' is for the quantile-t with one-sided (i.e. 1-alpha)
>PoC.Contrast=sum(Contrast.Coef*EstMean.aov)/(aov.sigma*
                  sqrt(sum(Contrast.Coef^2/n)))>
                  qt(1-alpha,df=(n-1)*len.Dosage)
# print the conclusion whether the PoC is established
>PoC.Contrast

TRUE
```

This established the PoC for these data and the next step is to identify the dose from the dose range where the statistically significant difference occurs with MCP, Mod, and MCP-Mod approaches.

9.4 Multiple Comparison Procedure (MCP) Approach

By nature of design, dose-response trials typically include multiple dose groups plus a control. Therefore a multiple comparison procedure should be utilized to adjust for multiplicity. For example, in this data, there are four doses of 0.05, 0.2,

0.6, 1 plus the placebo dose. There are many procedures in multiple comparison adjustment and we illustrate the most commonly used procedures, such as the Fisher's protected least significant difference (LSD), Bonferroni procedure, Dunnett's procedure, Holm's procedure, Hochberg, and gate-keeping procedures.

9.4.1 Fisher's Protected LSD (Fixed Sequence Test)

As a comparison, we first illustrate the naïve MCP procedure which was termed as Fisher's protected LSD (readers may refer to Sect. 3.1). This procedure is also known as the "fixed sequence test" in Westfall and Krishen (2012) and in Bretz et al. (2005) to determine the dose using a 5% one-sided level. This can be easily conducted using the R function "*lm*" for all pairwise comparison to placebo as follows:

```
# pairwise comparison to control without adjustment using ''lm''
>fit.aov2Cont =lm(resp~as.factor(dose), biom)
# print the model fit
>summary(fit.aov2Cont)
```

In the above R code chunk, the "lm" is used as an R function for one-way analysis of variance with "dose" as a "factor" and the entire model is an R object named as "`fit.aov2Cont`" to be used for future analysis. With this object, we can call R function "summary" to reproduce the results reported in Table 9.2 from Bretz et al. (2005) as follows:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.3449     0.1593  2.165 0.03287
as.factor(dose)0.05   0.1118     0.2253  0.497 0.62068
as.factor(dose)0.2    0.4654     0.2253  2.066 0.04155
as.factor(dose)0.6    0.5895     0.2253  2.617 0.01032
as.factor(dose)1      0.6038     0.2253  2.680 0.00867

Residual standard error: 0.7124 on 95 degrees of freedom
Multiple R-squared:  0.1153,   Adjusted R-squared:  0.07808
F-statistic: 3.096 on 4 and 95 DF, p-value: 0.01921
```

It can be seen that the overall model fitting is statistically significant with two-sided p-value = 0.019 from the F-test statistic of 3.096 with degrees of freedom of 4 and 95. The values in the column "Estimate" corresponding to "dose" 0.05, 0.2, 0.6 and 1 are the estimated mean difference to the placebo dose, that is, $\mu_i - \mu_0$, which are 0.1118, 0.4654, 0.5895 and 0.6038, respectively. Examining the corresponding p-values at the column "Pr(>|t)", we can see that they are all

statistically significant as two-sided t-test except the dose at 0.05. As a default setting in R function "lm," the column "Pr(>|t|)" reported the 2-sided p-values which can be converted a one-sided *p*-value. The one-sided raw *p*-values can then be calculated as follows:

```
# Convert the raw 2-sided p-values to 1-sided from the model fit:
#  to extract the p-values (i.e. ''Pr(>|t|)'') from the ANOVA
#  coefficients(i.e. ''coef'') without the intercept (i.e.,''-1'')
> raw.pval = summary(fit.aov2Cont)$coef[-1,''Pr(>|t|)'']/2
# print these p-values
>raw.pval

   (dose)0.05     (dose)0.2     (dose)0.6       (dose)1
 0.310339936 0.020774426  0.005160541   0.004334960
```

This reproduced the results at their Table 9.2 (page 743) from Bretz et al. (2005). Therefore, this fixed sequence test concluded that the top three doses (i.e. 0.2, 0.6 and 1) were statistically significantly different to placebo.

The model diagnostics can be done by plotting the R object "fit.aov2Cont" using R code "plot(fit.aov2Cont)" which can be used to examine the residuals with the residual plots and QQ-plot for residual normality. For these data, we do not show the plot since there is no obvious deviation from the residuals.

9.4.2 Bonferroni Correction

As a way to adjust for multiple comparisons, the classical Bonferroni correction (Bonferroni 1935, 1936) adjusts the raw *p*-values, ensuring strong family-wise error-rate (FWER) control under arbitrary dependence of the raw *p*-values. It simply multiplies each raw *p*-value by the total number of hypotheses. We make use of the R package of "mutoss" which is maintained by the MUTOSS (multiple hypotheses testing in an open software system) team and can be downloaded from http://mutoss.r-forge.r-project.org/. This package is designed to unify the application and comparison of multiple testing procedure. Most of the MCP procedures are implemented in this package and we only need to call the corresponding function. The Bonferroni correction procedure can be done as follows:

```
# load the ``mutoss'' library
> library(mutoss)
# call the ''Bonferroni'' function for Bonferroni correction:
```

```
#   with option ''silent=FALSE'' to generate output
> Bon.MCP = bonferroni(raw.pval, alpha, silent=FALSE)
```

In the above code, `silent = FALSE` is to generate output (another option is `silent = TRUE` which means no output is generated. The "`raw.pval`" is the vector of unadjusted p-values for the pairwise comparison (i.e., $\mu_i - \mu_0$) which is from the ANOVA model fitting named "`fit.aov2Cont`" and "`alpha`" is the overall one-sided type I error or level of significance at which the FWER shall be controlled, which is set at alpha (α) = 0.05. The output of this Bonferroni correction can be shown as follows:

```
> Bon.MCP
$adjPValues
(dose)0.05  (dose)0.2   (dose)0.6   (dose)1
1.00000000 0.08309770  0.02064217  0.01733984
$rejected
(dose)0.05  (dose)0.2  (dose)0.6   (dose)1
FALSE    FALSE    TRUE     TRUE
```

From this output, we can find the adjusted p-values associated with each comparison of the specific dose to placebo. With alpha (α) = 0.05, it also showed whether the comparison was rejected or not. Among the four pairwise multiple comparisons, the doses of 0.05 and 0.2 were not rejected, whereas the top two doses of 0.6 and 1 were rejected.

9.4.3 Dunnett's Test

Dunnett's test is also to compare each of the dose treatments to the placebo treatment. This test was developed in Dunnett (1955) and updated in Dunnett (1964). The procedure takes advantage of positive correlations among pairwise comparisons. The fact that each dose is compared with the same placebo control makes all pairwise comparisons to be positively correlated. Dunnett's test was widely used in multiple comparison procedure for simultaneously comparing, by interval estimation or hypothesis testing, all active dose treatments with placebo when the outcome is sampled from a distribution where the normality assumption is reasonable. This test is designed to hold the *familywise error rate* at or below α when performing multiple comparisons of treatment group with control. To implement the Dunnett's test, we make use of the R package "DoseFinding" which can be implemented as follows:

```
# load ''DoseFinding'' library
>library(DoseFinding)
```

```
# make Dunnett contrasts matrix with R row-bind function ''rbind''
> contMat <- rbind(-1, diag(4))
# name this matrix for its rows with ''rownames''
#   and columns with ''colnames'' as seen in the output below
> rownames(contMat) <- Dosage
> colnames(contMat) <- paste(''D'', Dosage[-1], sep='' '')
# ''MCTtest'' to calculate the adjusted p-value for Dunnett's test
> Dunnett.MCP = MCTtest(dose, resp, data=biom, contMat = contMat)
# print the result for Dunnett's test
> Dunnett.MCP
```

The "MCTtest" function gave the Dunnett's contrast to compare all the dose treatments to the placebo and the test-statistic with the adjusted *p*-values for each dose treatment as follows:

```
Multiple Contrast Test
Contrasts:
     D0.05 D0.2 D0.6 D1
0       -1    -1   -1  -1
0.05     1     0    0   0
0.2      0     1    0   0
0.6      0     0    1   0
1        0     0    0   1

Multiple Contrast Test:
      t-Stat   adj-p
D1    2.680 0.0153
D0.6  2.617 0.0178
D0.2  2.066 0.0653
D0.05  0.497 0.6033
```

It can be seen from the multiple contrast test that the doses of 0.05 and 0.2 were not statistically significant and the other top two doses of 0.6 and 1 were statistically significant (adjusted *p*-value less than the nominal alpha of 0.05).

9.4.4 Holm's Step-Down Procedure

The Holm's step-down-procedure was developed in Holm (1979) and detailed in Huang and Hsu (2007). The terms "step-down" or "step-up" are in reference to the absolute values of t-test. "Step-down" implies the procedure starts with the pairwise *t*-test with the largest absolute value, and then steps down to the smaller absolute

value of *t*. It controls the family-wise error rate (FWER) and offers a simple test uniformly more powerful than the Bonferroni correction. The Holm's procedure is implemented in the R package "mutoss" with function "holm." This function can be simply called to analyze the "biom" data as follows:

```
# call Holm's procedure from ''mutoss'' to adjust the raw p-value
> Holm.MCP = holm(raw.pval, alpha)
# print the result from Holm's procedure
> Holm.MCP
```

The function "holm" is called to adjust the raw *p*-values "raw.pval" to control FWER and the output is produced as follows:

```
$adjPValues
(dose)0.05    (dose)0.2    (dose)0.6    (dose)1
0.31033994   0.04154885  0.01733984  0.01733984
$rejected
(dose)0.05  (dose)0.2  (dose)0.6   (dose)1
FALSE     TRUE    TRUE      TRUE
```

The Holm's test showed that the dose of 0.05 was not rejected. However, the doses of 0.2, 0.6 and 1 were rejected as each of them is statistically significantly different from placebo.

9.4.5 Hochberg Step-Up Procedure

The Hochberg procedure (Hochberg 1988) is based on marginal *p*-values. It controls the FWER in the strong sense under joint null distributions of the test statistics that satisfy Simes' inequality as summarized in Sarkar and Chang (1997). Huang and Hsu (2007) compared this Hochberg procedure with Holm's procedure in Sect. 9.4.4 and concluded that the Hochberg procedure is more powerful than Holm's (1979) procedure, but the test statistics need to be independent or have a distribution with multivariate total positivity of order two or a scale mixture thereof for its validity (Sarkar and Chang 1997). Whereas Holm's procedure is a step-down version of the Bonferroni test, Hochberg's is a step-up version of the Bonferroni test. Note that Holm's method is based on the Bonferroni inequality and is valid regardless of the joint distribution of the test statistics.

The Hochberg procedure is implemented in R function "hochberg" in R package "mutoss". This function can be called to analyze the "biom" data as follows:

```
# call ''hochberg'' to adjust the raw p-values
> Hoch.MCP <- hochberg(raw.pval, alpha)
# print the result from procedure
> Hoch.MCP
```

The output from this Hochberg procedure is as follows:

```
$adjPValues
(dose)0.05    (dose)0.2    (dose)0.6    (dose)1
0.31033994   0.04154885   0.01548162   0.01548162
$criticalValues
[1] 0.01250000 0.01666667 0.02500000 0.05000000
$rejected
(dose)0.05  (dose)0.2  (dose)0.6  (dose)1
FALSE     TRUE     TRUE     TRUE
```

With the same conclusion reached as the Holm's procedure, the Hochberg's procedure showed that the dose of 0.05 was not rejected and the doses of 0.2, 0.6 and 1 were rejected and hence statistically significantly different from placebo.

9.4.6 Gate-Keeping Procedure

Gate-keeping procedures are proposed to test hierarchically ordered hypotheses from clinical trials. The hypotheses are grouped into ranked families, such that the hypotheses in higher order ranked families are tested first and the lower ranked families can be tested only if rejections achieved for higher ranked families. In other words, the higher ranked families serves as gatekeepers for lower ranked ones.

In general, there are two types of gatekeeping procedures known as serial and parallel gate-keeping procedure. As discussed in Bauer et al. (1998) and Dmitrienko et al. (2006), the serial gatekeeping requires all hypotheses in one family be rejected before testing the next family, whereas parallel gatekeeping as discussed in Dmitrienko et al. (2003) only requires at least one hypothesis be significant in order to test the next one. The gatekeeping approaches were first proposed as closed tests as discussed in Marcus, Peritz and Gabriel (1976) by considering all possible configurations of the null hypotheses where in general, the number of computational steps grows exponentially with n (about 2^n calculations need to be performed to test n null hypotheses). Under the assumption of the nondecreasing dose response with respect to increasing doses, the multiple testing of individual doses could fit into the serial gate-keeping procedure by ordering the testing sequence from the highest to the lowest dose. A lower dose is tested only if all preceding higher doses are tested and rejected at level of significance (alpha, α). The testing stops once a hypothesis fails to be rejected.

This gate-keeping procedure can be easily implemented to start with the highest dose and then goes down to find the minimal dosage where dosage to placebo is statistically significant at this dosage, all higher doses should be significant. The R implementation can be done as follows:

```
# with the ranked dosages and the raw p-values
>GK.MCP = raw.pval < alpha
>GK.MCP
(dose)0.05 (dose)0.2 (dose)0.6 (dose)1
FALSE    TRUE    TRUE    TRUE
```

Same as the Holm's and Hochberg's procedures, the gate-keeping procedure showed that the dose of 0.05 was not rejected and the doses of 0.2, 0.6 and 1 were rejected, with each being statistically significantly different from placebo at the 0.05 level of significance.

9.5 Modeling Approach (Mod)

Different form the MCP approach discussed in the Sect. 9.4, the modeling approach (Mod) treats dose as a continuous variable and assumes a functional relationship between the response and the dose. Sometimes, the multiplicity issue is avoided as a result of the assumed functional relationship (e.g., non-decreasing dose–response relationship).

9.5.1 Dose-Response Models

The dose-response models are used to describe certain functional relationship between the observed responses y_{ij} and the corresponding dose d_i, where i indicates the dose-level from 1 to k ($k = 5$ in the "biom" data) and j indicates the observations in each dose level from 1 to n_i ($n_i = 20$ in the "biom" data for all 5 dosages). This model can be mathematically expressed as follows:

$$y_{ij} = f(d_i, \theta) + \varepsilon_{ij}$$

where ε_{ij} is the error term. The mean response $\mu_i = E(y_{ij})$ and the corresponding dose d_i can then be expressed as

$$\mu_i = f(d_i, \theta)$$

where $f(.)$ is a linear or nonlinear function of dose d and the vector of modeling parameters θ. There are potentially many types of dose-response models, such as the linear model, linear in log-dose model, the exponential model, the quadratic

model, the logistic model, the Emax model, the sigmoid Emax model and the beta model. These models are briefly summarized in Table 9.2 with their functional form and the associated model parameters.

The graphical presentation of these models can be shown using the following R code chunk which produced Fig. 9.2.

```
# make a model list
modlist = list(linear = NULL,linlog=1, emax = c(0.05, 0.2),
 logistic = matrix(c(0.25,0.7, 0.09,0.06), byrow=FALSE, nrow=2),
 exponential = 1, sigEmax=c(0.1,1), quadratic=-1)
 # call; ''plotModels''; function; to; plot; these; models
 plotModels(modlist, Dosage, base=0, maxEff=1)
```

In this presentation, we specified these models with some selected parameters and also we illustrated two Emax models and two logistic models. As seen from Fig. 9.2, all these models except for the quadratic model are monotonic increasing as doses increase.

Among these models, a very popular dose-response model is the sigmoid Emax model as seen in Table 9.2. This sigmoid Emax model is a four-parameter model

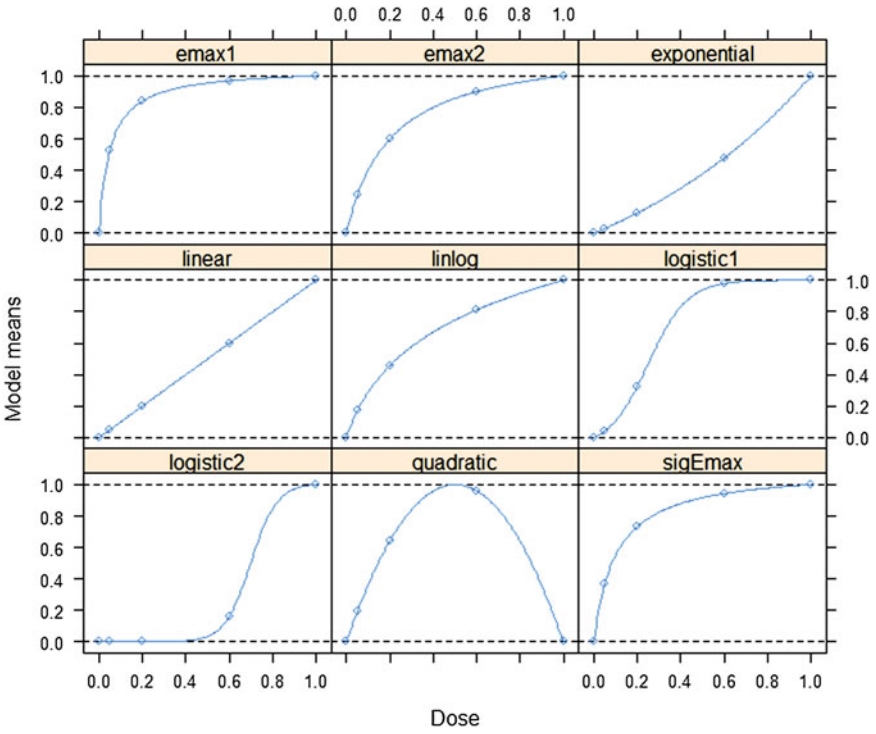

Fig. 9.2 Shapes of dose-response models in Table 9.2

where E_0 is the effect at dose 0 (the placebo response); *E*max is the maximum effect attributable to the drug; ED_{50} is the dose, which produces half of Emax; and λ is the slope factor (Hill factor). In practice, the three-parameter Emax model is typically considered, assuming the Hill factor $\lambda = 1$. As discussed in Thomas (2006), the Emax model can fit very well for most of the actual dose-response data and we illustrate this Emax model in this chapter for the Mod approach.

With the estimated Emax dose-response model, the doses of interest can be easily estimated from this model. Among the doses of interest, the ED_{50} is the most interested dosage to be estimated. It is in fact embedded in the Emax model and can be obtained from the model fitting. Other doses of interest, such as ED_{90}, the dose that produces 90% of maximum effect (E_{max}) from the baseline (E_0), which is sometimes used as an estimate of the maximum effect dose (MaxED), can be also estimated. In general, for any p%, the associated dosage of interest from Emax model can be derived as:

$$ED_p = ED_{50}\left(\frac{p}{1-p}\right) \tag{9.3}$$

With this formulation, we these that $ED_{20} = 0.25 \times ED_{50}$ and $ED_{90} = 9 \times ED_{50}$.

9.5.2 R Step-by-Step Implementations

We fit the Emax dose-response model to the "biom" data using the R function "`fitMod`" as follows:

```
# fit the ''Emax'' model
>fitemax = fitMod(dose, resp, data=biom, model=''emax'')
```

With the Emax model, the model fitting can be graphically illustrated in Fig. 9.3 by calling R function "`plot`" to the fitted model object "`fitemax`" above as follows:

```
# Call ''plot'' to plot the Emax fit
>plot(fitemax,CI=TRUE,plotData=''meansCI'',
   xlab=''Dose'',ylab=''Response'')''
```

We used the option "CI = TRUE" to add the confidence interval for the fitting. The option plotData = "meansCI" to add the dashed vertical lines for the confidence intervals for the means. It can be seen from Fig. 9.3 that the Emax model fitted the "biom" data very satisfactorily. In Fig. 9.3, the dashed vertical lines at each observed dosages indicated the confidence intervals for the means at those

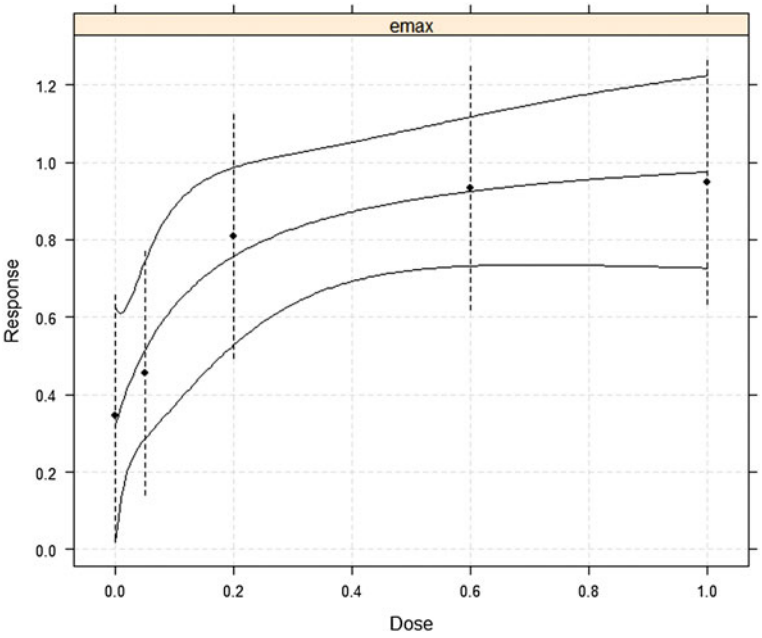

Fig. 9.3 Model fitting to "biom" using Emax model where the curves as upper 95% CI, predicted, lower 95% (*from top to bottom*)

observed dosages. The solid curve in the middle is the fitted Emax model, which is closed to the observed mean responses at each observed dosages. The confidence interval for the fitted Emax model can be seen from the upper and lower curved lines about the fitted Emax model.

With the established PoC from previous studies, we go further to compare the dosages against placebo and then test for differences them.. The R implementation is as follows:

```
# call ''predict'' to estimate the difference
>EstEff = predict(fitemax, se.fit=TRUE,
              doseSeq=Dosage, predType=c(''effect''))
# print the estimated differences and associated SE
>EstEff
# calculate the t-statistics and p-values
> t.estEff = EstEff$fit/EstEff$se.fit
# compute the p-values
> pval.estEff = 1-pt(t.estEff[-1], df=fitemax$df)
# print the p-values
> pval.estEff
```

In the above R code chunk, we used the R function "predict" with option predType = c("effect") at the observed dosage levels at "Dosage", which would output the estimated effects (i.e., the estimated mean differences at each dose-level relative to the placebo). We used the options "se.fit = TRUE" to output the standard errors for these differences. Doing so gave the following estimated mean differences to placebo and associated standard errors at the 5 dosage levels in the "biom" data:

```
$fit
0.0000000 0.1941593 0.4361932 0.6033237 0.6533942

$se.fit
0.0000000 0.1654233 0.2111702 0.1837320 0.1868323
```

These values are then used to calculate the *t*-statistics and the associated *p*-values at dose-levels of 0.05, 0.2, 0.6 and 1 by using the R function "pt":

```
(dose)0.05     (dose)0.2      (dose)0.6      (dose)1
0.1216920183  0.0207664633  0.0007128031  0.0003555304
```

It can be seen that treatment effects start from dose 0.20 and indicates that the top three doses of 0.2, 0.6 and 1 are all statistically different from placebo.

Using this estimated Emax dose-response model, we can easily estimate the doses of interest along with their confidence intervals. For example, the estimation of ED_{50} can be done by using the following step-by-step R code chunk:

```
 # extract the estimated 3 parameters from Emax model fit
>est.parm = coef(fitemax)
# extract the SE for the 3 estimated parameters
>se.parm  = diag(sqrt(vcov(fitemax)))
# estract the estimated ED50
> ED50 = est.parm["ed50"]
> ED50

   ed50
0.1421871

# extract the SE for ED50
> se.ED50 = se.parm["ed50"]
> se.ED50

0.1804621
```

```
# To construct 95% CI, extract the df from Emax model
> df.emax = fitemax$df
> df.emax

[1] 97

# calculate the 95% quantile
> q.975 = qt(0.975, df.emax)
> q.975

[1] 1.984723

# the lower bound for 95% CI
> low.ED50 = ED50-q.975*se.ED50
> low.ED50

-0.2159802

# the upper bound for 95% CI
> up.ED50 = ED50+q.975*se.ED50
> up.ED50

0.5003544
```

From this above illustration, we can see that the estimated $ED_{50} = 0.142$ has a relative large standard error of 0.180. Since the degrees of freedom is 97 from the Emax model, the 95% upper quantile would be 1.985. Therefore the with 95% CI for the estimated ED_{50} can be calculated as (−0.216, 0.500). Since the dose cannot be zero, we would use (0, 0.500) as the adjusted 95% confidence interval for ED_{50}.

For any other dosages of interest, we can use the Eq. (9.3) with the estimated ED_{50} and its standard error. This can be illustrated with the following R code chunk:

```
#calculate EDp for p from 0.1 to 0.95 by 0.05
p=seq(0.1,0.95, by=0.05)
# use equation 9.3 to calculate EDp
EDp = ED50*p/(1-p)
# the se for EDp
se.EDp = se.ED50*p/(1-p)
# the lower 95% CI
low.EDp = EDp-q.975*se.EDp
```

```
# the upper 95% CI
up.EDp  = EDp+q.975*se.EDp
# adjust the EDp and its CI between the 0 and 1
EDp.adj = pmin(1,ED50*p/(1-p))
low.EDp.adj = pmax(0,EDp-q.975*se.EDp)
up.EDp.adj  = pmin(1,EDp+q.975*se.EDp)
```

This calculation can be summarized in Table 9.3. As seen in Table 9.3, we calculated the EDs for percentages from 10 to 95% by 5%. In Table 9.3, we reported two panels. The middle panel reported the EDs and its 95% CI without adjustment. For this dose-response data, the estimated standard error for ED50 is relative large and yielded the lower CI bound to be negative, which was adjusted to be zero to corresponding to the original dose-response design from 0 to 1 (see right panel). Similarly, there are estimates for the upper CI bounds of EDp, which were greater than 1; those values were also adjusted, this time to 1 as shown in the right panel.

This result can be graphically illustrated in Fig. 9.4. In Fig. 9.4, the line linked the dots in the middle is the adjusted estimated EDs and the dashed lines at the top and the bottom represent the adjusted 95% CI bounds. Note that the lower 95% CI line is horizontal at fixed value of 0.

9.6 MCP-Mod Approach

9.6.1 Introduction

As the third approach in analyzing dose-response trials, the MCP-Mod method is to combine the MCP approach in Sect. 9.4 and Mod approach in Sect. 9.5 into one unified approach by first selecting the dose–response model using MCP to protect the FWER; then, the selected model is used to estimate the dose-response relationship. This approach is flexible in the sense of model selection; on the other hand, it pays the price of splitting α at the model selection step.

This MCP-Mod approach was developed by Bretz et al. (2005) and they further developed an R package (i.e. "MCPMod") (Bornkamp et al. 2009) which is available at http://cran.r-project.org/web/packages/MCPMod/index.html. We do not intend to replicate the theory in this chapter but instead concentrate on illustration of this package for real data analysis. We again make use of the "biom" data for this illustration. This illustration is largely described in Bornkamp et al. (2009) and this chapter is a further application of this package.

Table 9.3 Estimation of effective-doses at different percentages of p (EDp) and their 95% CI

	Un-adjusted ED_p estimation			Adjusted ED_p estimation		
p	EDp	Lower CI	Upper CI	EDp	Lower CI	Upper CI
0.10	0.016	−0.024	0.056	0.016	0	0.056
0.15	0.025	−0.038	0.088	0.025	0	0.088
0.20	0.036	−0.054	0.125	0.036	0	0.125
0.25	0.047	−0.072	0.167	0.047	0	0.167
0.30	0.061	−0.093	0.214	0.061	0	0.214
0.35	0.077	−0.116	0.269	0.077	0	0.269
0.40	0.095	−0.144	0.334	0.095	0	0.334
0.45	0.116	−0.177	0.409	0.116	0	0.409
0.50	0.142	−0.216	0.5	0.142	0	0.5
0.55	0.174	−0.264	0.612	0.174	0	0.612
0.60	0.213	−0.324	0.751	0.213	0	0.751
0.65	0.264	−0.401	0.929	0.264	0	0.929
0.70	0.332	−0.504	1.167	0.332	0	1
0.75	0.427	−0.648	1.501	0.427	0	1
0.80	0.569	−0.864	2.001	0.569	0	1
0.85	0.806	−1.224	2.835	0.806	0	1
0.90	1.28	−1.944	4.503	1	0	1
0.95	2.702	−4.104	9.507	1	0	1

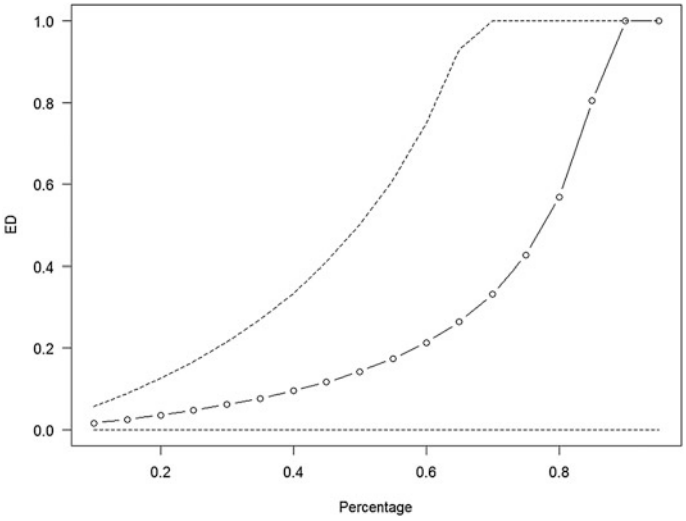

Fig. 9.4 The estimated EDs at different percentages and the associated 95% CIs

9.6.2 Step-by-Step Implementations in R Package "MCPMod"

The first step to use MCPMod is to select a candidate set of models for approximating the underlying dose-response relationship. The package gave a list of candidate models to be used, such as linear model, log-linear model, quadratic model, logistic model, exponential model, beta-model, Emax model, and sigmoid Emax model. In addition to these models, the users can also define their own non-linear models. The candidate models should be specified as a list with the list elements named according to the underlying dose-response model function, along with guesstimates or NULL for the model parameters. For illustration purpose, let's specify the candidate model lists for the "biom" dose-response data as follows:

```
# Specify the candidate model list
> modlist = list(linear=NULL, linlog=NULL,
          betaMod = c(0.5, 1), logistic=c(0.25, 0.09),
          quadratic=-1, exponential=1,
          emax=c(0.05,0.2), sigEmax=c(0.1,1))
# Plot the model list
>plotModels(modlist, doses=Dosage, base = 0,
          maxEff=1,scal = 1.2)
```

We deliberately select the candidate model list, which is different from the models used in Bretz et al. (2005), to show flexibility of the `MCPMod` to perform in model fitting and model selection. Bretz et al. (2005) used linear model, log-linear model, two exponential models, two quadratic models and the Emax model, whereas we select the linear model, log-linear model, beta model, logistic model, quadratic model, exponential model, Emax model and sigmoid Emax model, respectively, as seen in the above R code chunk.

We named the selected model list as an R object "`modlist`" and the numerical values are the guestimated parameter initial values. This model list can be plotted using the R function "plotModels" to be examined for different varieties of model shapes and distributions. Figure 9.5 gives the plot for this candidate model list. It can be seen that this candidate model list includes some monotonic increasing functions (such as the linear model, log-linear model, logistic model, exponential model, Emax model and sigmoid Emax model) as well as other shapes of models (such as beta model and quadratic model).

The main function in this package is also named as "`MCPMod.`" It provides all the functionalities of the full MCP-Mod approach and includes two key steps:

(1) MCP-step to calculate the optimal contrasts, critical value, contrast test statistics and possibly *p*-values, and selection of the set of significant models,
(2) Modeling step for model fitting, model selection/model averaging, and dose estimation.

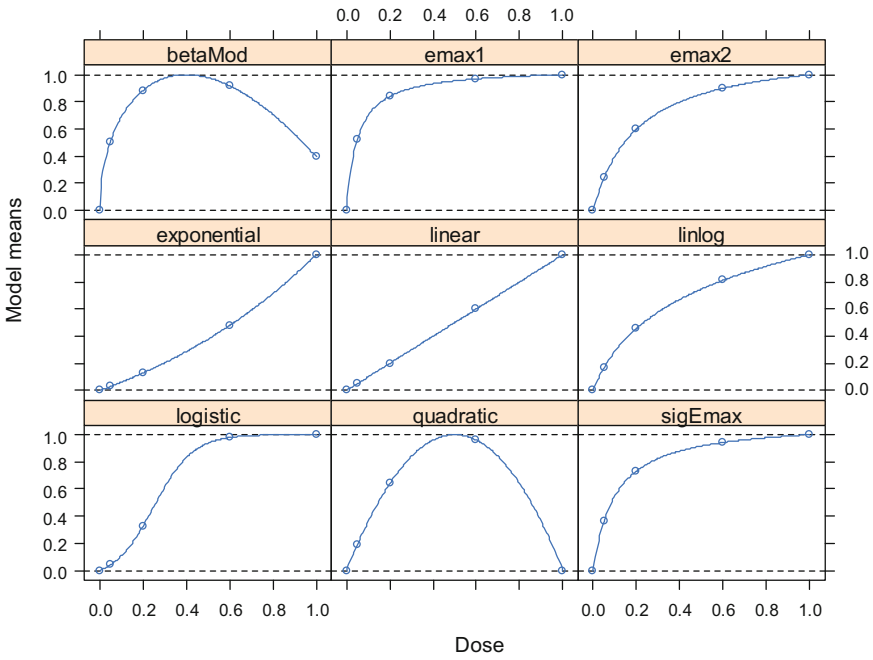

Fig. 9.5 Shapes of the candidate dose-response models

We call this function to analyze the "`biom`" and the following simple R code chunk is all that is needed for MCP-Mod analysis:

```
# Fit the MCP-Mod
>fit.MCPMod = MCPMod(biom, modlist, alpha = 0.05, pVal = TRUE,
 selModel = ''maxT'', doseEst = ''MED2'', clinRel = 0.4, dePar = 0.05)
# Print the summary of model fitting
>summary(fit.MCPMod)
```

We now explain some of the most important arguments used for the `MCPMod` function. For a complete description of the `MCPMod` function, we refer to the online documentation or use R "help" function as `help(MCPMod)`. It can be seen from the above R code chunk that the first argument is the dose-response data set (i.e., `biom` in this case) followed by the candidate model list (i.e., `modlist` defined previously). The `alpha` is the level of significance for the multiple contrast test, specified as `alpha = 0.05`. The `pVal` argument is to determine whether multiplicity adjusted *p*-value for the multiple contrast test should be calculated or not and we specified it as `pVal = TRUE`. The `selModel` argument is a dose-estimation criterion to determine how to select a dose-estimation model from the set of

significant models (if there are any significant models). Possible selections include the following:

- "MaxT" to select the model corresponding to the largest *t*-statistic (i.e., the maximum contrast test statistic which is the default setting),
- "AIC" to select the model with smallest AIC,
- "BIC" to select the model with smallest BIC, or
- "aveAIC" and "aveBIC" to select the model using a weighted average of the models corresponding to the significant contrasts (see Buckland et al. 1997 for details).

Among the four selections, the first three are specifically for each dose-response models whereas the fourth selection is the so-called "model averaging," which averages AIC/BIC from all the fitted dose-response models based on the list of model fitting with a weighted-average. As detailed in Buckland et al. (1997), the model weights for model k (k = 1, …, K) are calculated as $w_k = \frac{\exp(-0.5AIC_k)}{\sum_{k=1}^{K}\exp(-0.5AIC_k)}$ for AICs. The same formula can be modified for BICs.

The argument `doseEst` is to determine the minimum effective dose (MinED) once the overall dose-response relationship is established from an adequate model fitting. The MinED is defined as the smallest dose that demonstrates a clinically relevant and a statistically significant effect (denoted by Δ), specified as "`clinRel.`" Four dose-estimation options can be selected. One option is called "ED" to estimate the dose that gives a pre-specified percentage of the maximum effect. Specifically, the target dose ED_p is defined as the smallest dose that gives a certain percentage p of the maximum effect Δ observed in the dose range of $(d_1, d_k]$. The other three options are defined in Bretz et al. (2005) as "MED1," "MED2" and "MED3" with default selection of "MED2."

The `dePar` argument is used to specify additional parameters for the four dose estimators. For ED-type estimators, it determines which effective dose is estimated with the default 0.5 to estimate ED50. For MED-type estimators, it determines the confidence level γ used in the estimator. The used confidence level is given by 1 – 2 * `dePar` in corresponding to $1 - 2\gamma$. In the R code chunk, we specified `doseEst = ''MED2''` with `dePar = 0.05` and `clinRel = 0.4` to use the MED2 estimator with γ = 0.05 and the clinical threshold Δ = 0.4 to estimate the MED. This configuration means that we wish to determine the smallest dose such that the lower limit of the 95% confidence interval for the predicted response is greater than the predicted response for placebo and the point estimate is 0.4 above the predicted placebo level.

With all those settings, the summary of the MCPMod model fitting is given by `print(fit.MCPMod)` as follows:

```
Input parameters:
 alpha = 0.05 (one-sided)
 model selection: maxT
```

```
clinical relevance = 0.4
dose estimator: MED2 (gamma = 0.05)

 Optimal Contrasts:
     Lin    linlog  emax1 emax2 betaMod logistic   exp  sigEmax  quad
0    -0.437 -0.580 -0.799 -0.643  -0.714    -0.478  0.388 -0.723    -0.420
0.05 -0.378 -0.379 -0.170 -0.361  -0.043     0.435 -0.353 -0.287    -0.197
0.2  -0.201 -0.035  0.207  0.061   0.452    -0.147 -0.236  0.148     0.331
0.6   0.271  0.385  0.362  0.413   0.498     0.519  0.179  0.397     0.706
1     0.743  0.609  0.399  0.530  -0.192     0.540  0.798  0.465    -0.420

 Contrast  Correlation:
          linear linlog  emax1 emax2 betaMod logistic  exp  sigEmax quad
linear    1.000 0.961 0.766 0.912   0.229   0.945  0.992  0.848  0.071
linlog    0.961 1.000 0.903 0.990   0.489   0.976  0.922  0.958  0.323
emax1     0.766 0.903 1.000 0.949   0.774   0.828  0.705  0.986  0.526
emax2     0.912 0.990 0.949 1.000   0.606   0.956  0.860  0.988  0.431
betaMod   0.229 0.489 0.774 0.606   1.000   0.448  0.122  0.703  0.890
logistic  0.945 0.976 0.828 0.956   0.448   1.000  0.898  0.905  0.377
exp       0.992 0.922 0.705 0.860   0.122   0.898  1.000  0.789 -0.054
sigEmax   0.848 0.958 0.986 0.988   0.703   0.905  0.789  1.000  0.494
quad      0.071 0.323 0.526 0.431   0.890   0.377 -0.054  0.494  1.000

 Multiple Contrast Test:
             Tvalue  pValue
emax2         3.464   0.001
sigEmax       3.462   0.001
linlog        3.364   0.002
emax1         3.339   0.003
logistic      3.235   0.003
linear        2.972   0.007
exponential   2.752   0.012
betaMod       2.402   0.030
quadratic     1.850   0.093

 Critical value: 2.156

 Selected for dose estimation:
  emax

 Parameter estimates:
 emax model:
   e0   eMax  ed50
```

```
0.322 0.746 0.142

Dose estimate
MED2,90%
 0.17
```

The summary output includes five major parts:

(1) Input parameters for some information about important input parameters that were used for MCPMod modelling.
(2) "`Optimal Contrasts`" and "`Contrast Correlation`" and "`Multiple Contrast Test`" for the estimated optimal contrasts and the contrast correlations together with the contrast test statistics associated with the candidate models specified in the modlist.
(3) "`Multiple Contrast Test`" and "`Critical value`" are the contrast test statistics, the multiplicity adjusted p-values from the multivariate t-distribution in Bretz et al. (2005), and the critical value associated with the candidate models specified in the modlist. It should be noted in this part that all contrast tests with an adjusted p-value less than 0.05 or equivalently with t-value > 2.18 (the critical value) can be declared as statistically significant, with FWER maintained at 5% level. Of these nine candidate models only the quadratic model is not statistically significant, which is then not used in the reference set.
(4) `''Selected for dose estimation''` and `''Parameter estimates''` for the best fitted dose-response model from the candidate models and its parameter estimates. For these data, the best model is found to be the "Emax" model with estimated parameters as $E_0 = 0.322$, $E_{max} = 0.746$ and $ED_{50} = 0.142$.
(5) "Dose estimate" for the target dose estimate from this MCP-Mod modeling, which is 0.17 based on the settings in this modeling and close to the estimate from MoD approach at Sect. 9.4.

The fitted model from this MCP-Mod can be graphically displayed with the plot function in R as:

```
>plot(fit.MCPMod, CI=TRUE, doseEst = TRUE, clinRel = TRUE)
```

In the above R code, we plot the fitted MCPMod model with 90% confidence interval (i.e., `CI = TRUE`) and display the estimated dosage (i.e., `doseEst = TRUE`) and specified clinical relevance threshold (i.e., `clinRel = TRUE`). Figure 9.6 is produced with this R code. In this plot, the mean is plotted for each dosage (i.e., `Group Means`) with the estimated dose-response model (the Emax model for these data) prediction (i.e., `Model Predictions`) with 90% CI (i.e., `0.9 Pointwise CI`). The estimated MED (i.e., `Estim.`

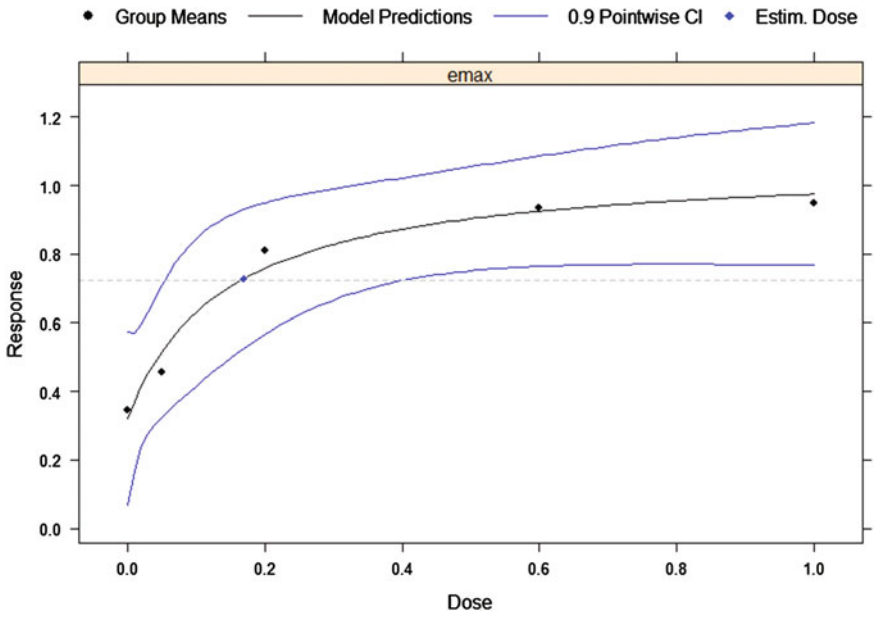

Fig. 9.6 Fitted Emax model with summary data

`Dose`) is at the intersection of the estimated dose-response model and the clinical relevance threshold (i.e., the dashed horizontal line).

In summary for the MCP-Mod approach, the best model selected for dose estimation from the model list is the Emax model and the estimated significant dosage is 0.17.

9.7 Discussion

In this chapter, we used a real data "`biom`" to illustrate the three approaches in analysis of dose- ranging trials, namely, the MCP approach, the Mod approach, and the combined MCP-Mod approach. We illustrated these procedures in a step-by-step implementation in R so that the interested readers can follow and understand the logic of these procedures and also use the R program for their own dose-response clinical trials

With the established PoC using the trend test in Sect. 9.3, we introduced six MCP procedures in Sect. 9.4 to control the FWER to identify the dosage which are statistically significantly different from placebo. These six approaches are the Fisher's protected LSD (fixed sequence test), the Bonferroni test, the Dunnett test, the Holm's test, the Hochberg test, and gate-keeping test. Among the six MCP tests, the Bonferroni test and the Dunnett test identified that the top two doses (i.e., 0.6

and 1) were statistically significantly different from placebo and the rest of four tests (i.e., the fixed sequence test, the Holm's test, the Hochberg test and gatekeeping test) identified the top three doses (i.e., 0.2, 0.6 and 1) were statistically significantly different from placebo.

The Mod approach in Sect. 9.5 used the Emax model to estimate the dose-response relationships. With the estimated dose-response Emax model, we further illustrated the estimation of doses of interest along with their confidence intervals. We also investigated other models, such as the log-linear and the sigmoid Emax model (as an extension of Emax model to include another parameter) and found that both models gave the same conclusions (not shown here due to the space limitations).

The combined MCP-Mod approach in Sect. 9.6 selected the best model as the Emax model from a list of dose-response models that included the linear model, log-linear model, logistic model, exponential model, Emax model, the sigmoid Emax model, the beta model, and the quadratic model. With the best model being Emax, the statistically significant dose was identified as 0.17.

References

Bauer, P., Röhmel, J., Maurer, W., & Hothorn, L. (1998). Testing strategies in multidose experiments including active control. *Statistics in Medicine, 17,* 2133–2146.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni* (pp. 13–60). Rome, Italy.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8,* 3–62.

Bornkamp, B., Pinheiro, J. C., & Bretz, F. (2009). MCPMod: An R package for the design and analysis of dose-finding studies. *Journal of Statistical Software, 29,* 1–23.

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

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics, 53,* 603–618.

Dmitrienko, A., Offen, W., & Westfall, P. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine, 22,* 2387–2400.

Dmitrienko, A., Wiens, B., & Westfall, P. (2006). Fallback tests in dose–response clinical trials. *Journal of Biopharmaceutical Statistics, 16,* 745–755.

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

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics, 20,* 482–491.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Huang, Y., & Hsu, J. (2007). Hochberg's step-up method: cutting corners off Holm's step-down method. *Biometrika, 94,* 965–975.

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

Pinheiro, J. C., Bretz, F., & Branson, M. (2006). Analysis of dose-response studies—modeling approaches. In N. Ting (Ed.), *Dose finding in drug development* (pp. 146–171). New York: Springer.

Sarkar, S. K., & Chang, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association, 92,* 1601–1608.

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

Wang, X., & Ting, N. (2012). A proof-of-concept clinical trial design combined with dose-ranging exploration. *Pharmaceutical Statistics, 11*, 403–409.

Westfall, P., & Krishen, A. (2012). Optimally weighted, fixed sequence and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference, 99,* 25–40.

Chapter 10
Data Analysis of Dose-Ranging Trials for Binary Outcomes

10.1 Introduction

Continuing the data analysis for dose-ranging clinical trials from Chap. 9 for continuous outcomes, we illustrate the analysis for binary data in this chapter. We will use the same data from Chap. 9, but dichotomize the continuous response data into binary data simply for the pedagogical purpose of continuing from the previous chapter and illustrating the methods on analyzing dose-ranging clinical trial with binary data.

The dataset "biom" was reported in Bretz et al. (2005) and involves a randomized double-blind parallel group trial with a total of 100 patients being allocated to either placebo or one of four active doses coded as 0.05, 0.20, 0.60 and 1, with 20 per group. The response variable was assumed to be normally distributed with larger values for a better outcome. There are two columns in "biom" named "dose" for dose levels and "resp" for responses.

For illustration in this chapter, we dichotomize all responses from all dose levels at the global median for the illustration of analyzing dose-ranging clinical trials with binary data, which is commonly called "median-splitting" in dichotomization for instructional purposes and continuation from the previous chapter. In Sect. 10.2, we describe the process of dichotomization. The resultant binary data is then used for binary dose-response modeling in Sect. 10.3 with four methods, which are Pearson's chi-squared test for proportions, Cochran-Armitage test for trend, and two versions of logistics regression models to establish significant dose-response relationship. In Sect. 10.4, we cover multiple comparisons and four multiple comparison procedures, which include the Bonferroni adjustment procedure, Bonferroni-Holm procedure, Hochberg procedure and the gate-keeping procedure. We conclude this chapter with some discussions in Sect. 10.5. All the illustrations are implemented in R with a detailed step-by-step fashion so that interested readers can follow the models and the R program for their own dose-response modeling.

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_10

10.2 Data and Preliminary Analysis

For these data, the global mean and median for the continuous outcome are 0.699 and 0.545, respectively, which can be seen from Fig. 10.1. With skewed distribution, the sample median is more representative statistic to categorize the data and we then use it to dichotomize the responses. The interested readers can use the mean or any other values as cutoff to re-do the dose-response modelling and multiple comparison in this chapter following the R code illustrated in the chapter. We leave this as an exercise to interested readers.

Using the global median (0.545) as cutoff to dichotomize the response data with response greater than 0.545 to represent "better" efficacy, we can produce a binary dose-response data with 100 observations along with the observed continuous data as follows:

```
# Use median as cutoff for dichotomize the response
>cutoff = 0.545
# create a binary response as ''cresp''
```

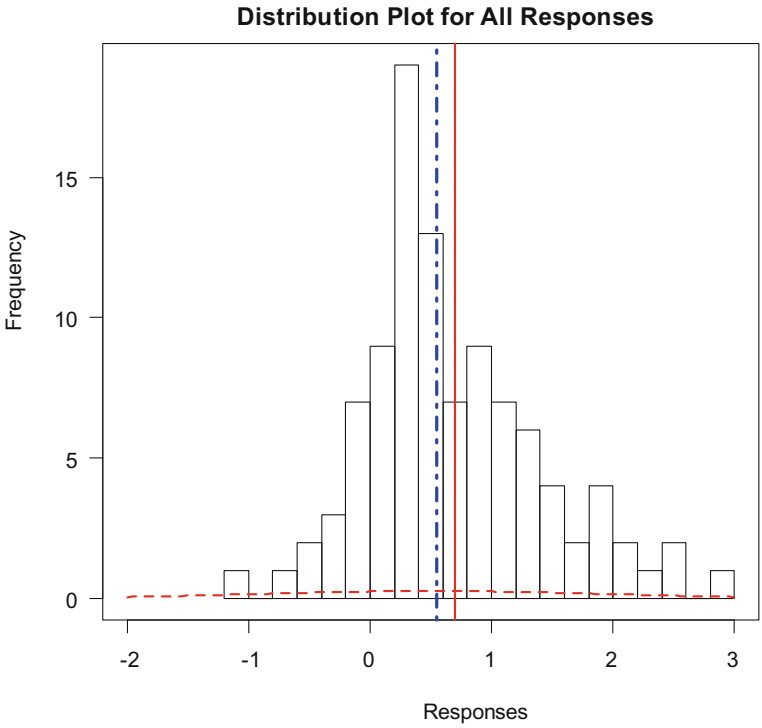

Fig. 10.1 Dose-response data distribution. The *vertical sold line* is the data mean and the *dashed vertical line* is the median. The *curved line* is the normal distribution using the data mean and standard deviation

```
>biom$cresp = ifelse(biom$resp > cutoff, 1,0)
# print the first 5 observations using ''head'' function
>head(biom, n = 5)

  dose        resp cresp
1 0.05  1.35380098   1
2 0.05  0.15498024   0
3 0.05 -0.07083804   0
4 0.05  0.58374091   1
5 0.05  0.96259853   1
```

In the above R code chunk, we employed the R function ifelse for this dichotomization at the cutoff = 0.545, which produced a binary response variable named as cresp. It can be seen from the above observations, "cresp" is 1 (i.e., "better") if "resp" is greater than the median of 0.545; otherwise the "cresp" is 0 (i.e., "worse"). A contingency table can be created to look at the binary data distribution using the following R code chunk:

```
# call R ''table'' function to create a summary data ''dat.tab''
>dat.tab = data.frame(table(biom$dose, biom$cresp))
# create the variable names for the table
>colnames(dat.tab) = c(''dose'',''better'',''cresp'')
# print the summary table
> dat.tab

   dose better cresp
1     0      0  14
2  0.05      0  14
3   0.2      0   9
4   0.6      0   7
5     1      0   6
6     0      1   6
7  0.05      1   6
8   0.2      1  11
9   0.6      1  13
10    1      1  14
```

We can see from this output that there are 50 (= 6 + 6 + 11 + 13 + 14) patients who are categorized into "better" category and 50 (= 14 + 14 + 9 + 7 + 6) who are categorized into "worse" category corresponding respectively to dosages 0, 0.05, 0.2, 0.6, and 1. The result of this dichotomization can be summarized into Table 10.1 and compared with Table 9.1 in Chap. 9.

This table can be then used to create a contingency table using the R function 'xtabs' as follows:

```
> out.dat = xtabs(cresp~dose+better, dat.tab)
> out.dat
     better
dose    0  1
  0    14  6
  0.05 14  6
  0.2   9 11
  0.6   7 13
  1     6 14
```

This contingency table can then be used to display the binary data distribution using the R function "`dotchart`" and "`mosaicplot`" as follows to produce Fig. 10.2:

Table 10.1 Binary dose-response data dichotomized from "biom"

Dose	Better (coded 1)	Worse (coded 0)
0.00	6	14
0.05	6	14
0.20	11	9
0.60	13	7
1.00	14	6

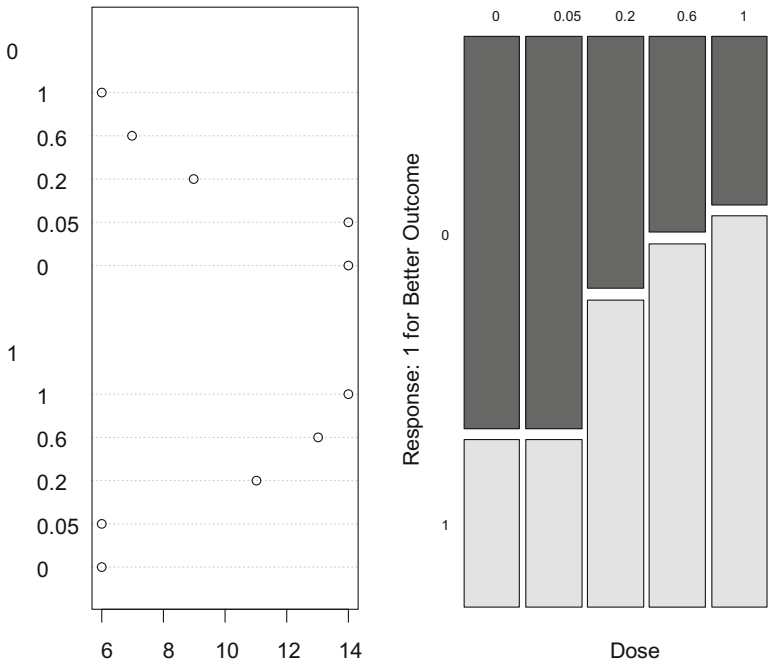

Fig. 10.2 The binary data distribution

```
# call R function ''dotchart'' to generate the dotplot
>dotchart(out.dat)
# call R function ''mosicplot''
>mosaicplot(out.dat,color=T,las=1,main ='''',xlab=''Dose'',
 ylab=''Response: 1 for Better Outcome'')
```

In Fig. 10.2, the left panel is the "dotchart" for the number of patients of binary response of 1 (i.e., "better") and 0 (i.e., "worse") from the total 20 patients at each dosage. The right panel is the corresponding "mosicplot" for the binary distribution. It can be seen from Fig. 10.2 that the proportions are the same for dosages 0 and 0.05 and then jump at dosages 0.2, 0.6 and 1. This figure is the graphical presentation for the binary data resulted from the dichotomization with 6, 6, 11, 13, and 14 patients in the 'better' category and 14, 14, 9, 7 and 6 in the "worse" category corresponding respectively to dosages 0, 0.05, 0.2, 0.6 and 1.

10.3 Modeling Approach

In modeling a dose-response relationship, we first test whether there is a statistically significant dose-response monotone relationship, with a more favorable response from lower to higher doses. Corresponding to these data, the monotonicity can be cast into a null hypothesis (H_0) of equal probability against the alternative hypothesis (H_1) of a monotone pattern among the five dose groups as formulated below:

$$H_0 : p_0 = p_{0.05} = p_{0.2} = p_{0.6} = p_1 \text{ versus } H_1 : p_0 \leq p_{0.05} \leq p_{0.2} \leq p_{0.6} \leq p_1 \tag{10.1}$$

At least one strict inequality from the above H_1 should be held true to ensure monotonicity. There are several methods which can be used for this hypothesis test and we demonstrate four commonly used methods in this section.

10.3.1 Pearson's χ^2-Test

The most commonly used method for binary responses is the Pearson's χ^2-test (chi-squared test) for categorical statistical analyses of contingency table data as described in Pearson (1900), Wilson (1927), Newcombe (1998a, b), and Agresti (2012). For this method, the first step in the test is to calculate the value of the chi-square statistic. It is obtained by (a) forming the difference between the observed number of frequencies and the expected number of frequencies (under the null hypothesis of no difference among the groups being compared) in each cell

of the contingency table, (b) squaring each difference, (c) dividing each squared difference by the expected number of frequencies, and (d) summing the results. The next step is to determine the degrees of freedom of the test, which is the total number of dosages less than 1 (i.e., 4 = 5 − 1 in these data).

The formula of the test statistic is

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{10.2}$$

where O_i is the observed frequency in the *i*th cell of the contingency table, E_i is the expected (theoretical) frequency under the null hypothesis of no difference in proportions, and the $\sum$ denotes summation. Asymptotically to the total number of counts, the distribution of the test statistic is a chi-square distribution. The "asymptotical" approximation to the chi-square distribution breaks down if expected frequencies are too low. In this case, a better approximation is obtained by using Yates' correction for lack of continuity. This is accomplished by reducing the absolute value of each difference between observed and expected frequencies by 0.5 before squaring. This χ^2test is implemented in R as function "`prop.test`." We can call this function to test this hypothesis as follows:

```
# ''table'' the original data
>dat = table(biom$dose, biom$cresp)
# make a data frame with the tabled data
Dat = data.frame(dose = c(0,0.05,0.2,0.6,1),
          worse=dat[,1],better=dat[,2],tot=dat[,1]+dat[,2])
# print the resultant data
> dat

   dose worse better tot
      0    14      6  20
   0.05    14      6  20
    0.2     9     11  20
    0.6     7     13  20
      1     6     14  20

# call ''prop.test'' for to test the hypothesis
prop.test(dat$better, dat$tot, alternative = c(''greater'') )
```

This "`prop.test`" can then produce the output as follows:

```
5-sample test for equality of proportions without continuity
     correction
```

```
data: dat$better out of dat$tot
X-squared = 11.6, df = 4, p-value = 0.02059
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3 prop 4 prop 5
  0.30   0.30   0.55   0.65   0.70
```

It can be seen from the output that the chi-square statistic is 11.6 with 4 degrees of freedom, which yields a statistically significant p-value of 0.021. From this test the null hypothesis of equal proportions is rejected in favor of the alternative that at least one proportion differs from another. In fact this can be seen from the estimated sample estimates of these five proportions as {0.30, 0.30, 0.55, 0.65, 0.70} show a monotonic increasing trend.

10.3.2 Cochran-Armitage Test for Trend

The second method is the well-known Cochran-Armitage test for trend named for William Cochran and Peter Armitage. As detailed in Cochran (1954) and Armitage (1955), the Cochran–Armitage test for trend is designed for categorical data analysis with the aim to assess the association between a variable with two categories and a variable with several categories. It modifies the Pearson chi-squared test described previously to incorporate a suspected ordering in the effects of these categories of the second variable, which can be used specifically for dose-response relationships using a binary outcome. Specifically to these data, there are doses of a treatment ordered from "low" to "high" in five categories as 0 (placebo), 0.05, 0.20, 0.60, and 1.

We hypothesize that the treatment benefit is monotonic and cannot become smaller as the dose increases. The Cochran-Armitage test is then designed to test the presence of this trend. Compared with the Pearson's chi-squared test, the trend test has higher power when the hypothesized trend is correct. This Cochran-Armitage test is implemented in the R package "`coin`" with a function "`independence_test`" as follows:

```
# load the ''coin'' package
> library(''coin'')
> perform Cochran-Armitage trend test for proportions
> independence_test(cresp~dose, data=biom, teststat=''quad'')
```

where "teststat=”quad" indicates that the type of test statistic is in quadratic form. The output of this test can be presented as follows:

```
        Asymptotic General Independence Test

data: cresp by dose (0 < 0.05 < 0.2 < 0.6 < 1)
chi-squared = 10.474, df = 1, p-value = 0.001211
```

It can be seen from the output that the associated p-value from this particular type of chi-squared test is 0.0012, which indicates that there is a significantly significant monotonic increasing dose-response trend.

10.3.3 Logistic Regression with Dose as Continuous Variable

The third method uses logistic regression as described in McCullagh (1980), Chen and Peace (2011) and Chen et al. (2017) with dose as a continuous variable (d_i, i = 1, …, 100 subjects) to the observed proportions (p_i) as follows:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 d_i. \tag{10.3}$$

In this model, the parameter β_0 is the intercept parameter to represent dose at placebo and β_1 is the slope parameter to be tested which represents the log odds-ratio for one-unit increase of dose. The maximum likelihood estimation can be used to estimate the model parameters and make statistical inferences, which can be performed by calling the R function "glm". In this model, the monotonic dose-response relationship can be characterized to test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 > 0$, so rejecting H_0: $\beta_1 = 0$ would imply that higher doses will produce more favorable response. Maximum likelihood estimation can be used to estimate the model parameters and make statistical inferences which can be performed by calling the R function "glm" as follows:

```
# call ``glm'' for logistic regression and name it as ``m3''
> m3 = glm(cbind(better,worse)~dose,
                family=binomial,data=dat)
# print the model fit
> summary(m3)
```

The summary of parameter estimates can be shown as:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.6222     0.2907  -2.141  0.03230
   Dose        1.7196     0.5801   2.964  0.00304
```

It can be seen that the estimated slope coefficient (β_1) is 1.719 (which is positive) with standard error of 0.58, producing a statistically significantly p-value of 0.003 and indicating a monotonic increasing responses.

Based the model (10.3), the estimated or predicted response rates are 0.349, 0.369, 0.431, 0.601 and 0.750 corresponding to doses of 0, 0.05, 0.2, 0.6 and 1, respectively, and compare with the observed response rates of 0.30, 0.30, 0.55, 0.65 and 0.70, as shown in Fig. 10.3. In Fig. 10.3, the dots are the observed response rates, the solid line represents the estimated (predicted) response rates, and the dashed lines represent the associated 95% confidence intervals for true response rates.

With the estimated parameters from the dose-response model in Eq. (10.3), we can easily estimate the doses of interest, such as effective dose of ED_{50} as the effective dose that produces 50% probability of success:

$$ED_{50} = -\frac{\widehat{\beta_0}}{\widehat{\beta_1}} \tag{10.4}$$

For any other effective doses ED_{p} at probability p, the formula in Eq. (10.4) can be extended as follows:

$$ED_p = \frac{logit(p) - \widehat{\beta_0}}{\widehat{\beta_1}} \tag{10.5}$$

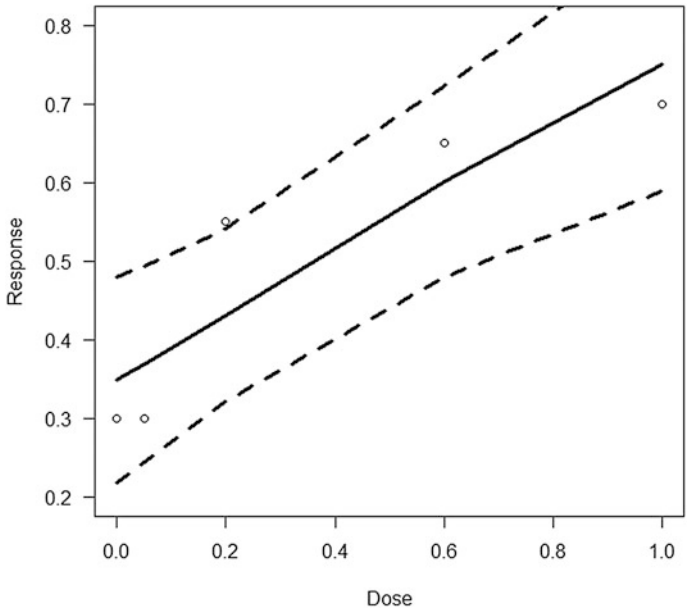

Fig. 10.3 Observed and estimated dose-response relationship

The standard errors for these effective doses in Eqs. (10.4) and (10.5) can be determined from the delta method which is to expand the Eq. (10.5) by Taylor series to the order of 1. This estimation is implemented in the R MASS library with the function dose.p. For illustration purposes, we can call this function to estimate the ED_{40}, ED_{50} and ED_{60} as follows:

```
# load the MASS library
> library(MASS)
# call ''dose.p'' function
> dose.p(m3, p=c(0.4,0.5,0.6))

             Dose       SE
p = 0.4: 0.1260292 0.1429848
p = 0.5: 0.3618175 0.1222877
p = 0.6: 0.5976058 0.1487271
```

Therefore the estimated ED_{40}, ED_{50} and ED_{60} are 0.126, 0.362 and 0.598, respectively, with corresponding standard errors of 0.143, 0.122 and 0.149.

10.3.4 Logistic Regression with Dose as Categorical Variable

The fourth method is to use logistic regression as follows:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_{0.00} + \beta_{0.05} + \beta_{0.20} + \beta_{0.60} + \beta_{1.00} \tag{10.6}$$

In this model, dose is treated as categorical so we estimate a parameter for each dose-level where $\beta_{0.00}$ represents the reference parameter corresponding to dose at placebo, and the rest of parameters of $\beta_{0.05}, \beta_{0.20}, \beta_{0.60}, \beta_{1.00}$ represent the differences from placebo. All these parameters are in fact log odds ratios. The maximum likelihood estimation can be used to estimate the model parameters and make statistical inferences, which can be performed by calling the R function "`glm`" and specifying the dose as factor in following R code chunk:

```
# call ''glm'' for logistic regression and name it as ''m2''
> m2 = glm(cbind(better,worse)~factor(dose),
                family=binomial,data=dat)
# print the model fit
> summary(m2)
```

The summary of parameter estimates can be shown as follows:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)- 8.473e-01  4.880e-01 -1.736  0.0825
(dose)0.05  5.417e-16  6.901e-01  0.000  1.0000
(dose)0.2   1.048e+00 6.634e-01  1.580  0.1142
(dose)0.6   1.466e+00 6.767e-01  2.167  0.0302
(dose)1     1.695e+00 6.901e-01  2.456  0.0141
```

It can be seen from the column of "`Pr(>|z|)`" that the two lower doses of 0.05 and 0.2 are not statistically significantly different from placebo, but the two higher doses of 0.6 and 1 are in this two-sided hypothesis test. These 2-sided p-values for testing $H_0 : p_i = p_0$ versus $H_1 : p_i \neq p_0$ can be converted to one-sided p-values, by dividing by 2, for testing $H_0 : p_i = p_0$ versus $H_1 : p_i > p_0$ as 0.5000, 0.0572, 0.0155 and 0.0070 for doses 0.05, 0.2, 0.6 and 1, respectively. The pattern of statistical significance for these pair-wise comparisons still holds in the one-sided hypothesis testing, which is not statistically significantly better than the placebo for the two lower doses of 0.05 and 0.2 but statistically significantly better for the two higher doses of 0.6 and 1. Based on this model, the estimated log odds are −0.847, −0.847, 0.201, 0.619 and 0.847 corresponding to doses of 0, 0.05, 0.2, 0.6 and 1, respectively, which again indicated monotonic increasing responses. Note that these one-sided p-values will be used in Sect. 10.4 for illustration of multiple comparisons.

10.4 Multiple Comparisons

10.4.1 The Raw p*-Values*

Section 10.3 covers the analysis when all treatment groups are included in one analysis. This current section will discuss how to handle pairwise comparisons of each dose against placebo to control using the family-wise error rate (FWER). With multiple doses are included in a dose-ranging trial, the goal of this section is to test the efficacy of each dose against the control (often a placebo) and the statistical analyses should be adjusted for multiple comparisons, focus of this section. It is now well-recognized that the primary objective of a multiple-testing procedure is to control the Type-I error rate (α) as the overall probability of erroneously rejecting at least one null hypothesis irrespective of which and how many of the null hypotheses of interest are in fact true.

For this purpose, the model in Sect. 10.3.4 can be employed for this multiple adjustment for pair-wise comparisons. As seen in Sect. 10.3.4, the estimated log odds (in column "Estimate") for placebo is −0.847, which is not statistically significant (p-value = 0.0825). The log odds1 ratios to for the rest of the dosages of 0.05, 0.2, 0.6 and 1 (relevant to placebo) are respectively 0.000, 1.048, 1.466 and 1.695 (as shown in column "Estimate"), with the associated standard errors of 0.690, 0.663, 0.677 and 0.690 (as seen in column "Std. Error"), which resulted in Wald z statistics of 0.000, 1.580, 2.167 and 2.456 (as seen in column "z value"). The two-sided p-values for testing $H_0 : p_i = p_0$ versus $H_1 : p_i \neq p_0$ on the log odds scale are 1.000, 0.114, 0.030 and 0.014 (as seen in column 'Pr(>|z|)"). The associated one-sided p-values for testing $H_0 : p_i = p_0$ versus $H_1 : p_i > p_0$ on the log-odds scale become 0.5000, 0.0571, 0.0151 and 0.0070 for doses 0.05, 0.2, 0.6 and 1, respectively.

To control Type 1 error rate for these tests, we focus on the MCP approaches to adjust p-values significance levels from these one-sided p-values for comparing each dose to the placebo. We will assume throughout in Sect. 10.4 that the overall Type 1 error rate is to be controlled at the 5% level. We can extract these p-values from the logistic regression in Sect. 10.3.4 as follows:

```
# unadjusted p-values on pair-wise comparison
# use ''-1'' to remove the associated p-value for placebo
> raw.pval = summary(m2)$coef[-1,''Pr(>|z|)'']/2
# print the raw p-values
> raw.pval

(dose)0.05     (dose)0.2     (dose)0.6     (dose)1
0.500000000  0.057091821 0.015117195 0.007030475
```

10.4.2 Bonferroni Adjustment

Because of its simplicity, Bonferroni adjustment is often used despite its conservativeness. Under this approach, the overall Type I error rate is adjusted to the number of tests. That is, in the current example, we will compare each raw p-value to 0.0125 (= 0.05/4) since there are 4 comparisons. Comparing the unadjusted raw (observed) p-values to 0.0125, we can see that only the p-value associated with dose of 1 is smaller than 0.0125. Thus, applying the Bonferroni procedure, we can conclude only that the highest dose (i.e., dose = 1) produces a significantly better result than the placebo.

Using the R mutoss library, we can call R function bonferroni as follows:

```
# load R "mutoss" library
> library(mutoss)
> alpha =0.05
# call "Bonferroni" function
> Bon.MCP =  bonferroni(raw.pval, alpha, silent=FALSE)
# print it
>Bon.MCP

                Bonferroni correction

$adjPValues
(dose)0.05  (dose)0.2   (dose)0.6    (dose)1
1.00000000  0.22836728  0.06046878   0.02812190

$rejected
(dose)0.05  (dose)0.2  (dose)0.6    (dose)1
     FALSE      FALSE      FALSE       TRUE

$errorControl
An object of class "ErrorControl"
Slot "type":
[1] "FWER"
```

Some explanations are needed for this R implementation. Instead of adjusting the FWER (α) to 0.0125, the R bonferroni function expands the unadjusted raw p-values (i.e., 0.500, 0.057, 0.015, 0.007) by the number of comparisons annotated as \$adjPValues in the output as 1.000, 0.228, 0.060, and 0.028 and compares these adjusted p-values to the overall FWER ($\alpha = 0.05$), where only the p-value at dose 1 is greater than 0.05. Therefore the bonferroni function outputs which comparison is rejected as annotated as \$rejected to be TRUE at dose = 1. The entire procedure are summarized as follows:

```
Number of hyp.: 4
Number of rej.:  1
 rejected     pValues adjPValues
1       4 0.007030475 0.0281219
```

10.4.3 Bonferroni–Holm Procedure

The Bonferroni-Holm procedure begins with ordering the unadjusted p-values from the smallest to the largest and then compares these ordered p-values with a significance level that is FWER (at $\alpha = 0.05$)] divided by the number of hypotheses remaining to be tested at each stage until the p-value under comparison becomes larger than the current significance level. When this occurs, the procedure would stop and conclude significance for all comparisons before the present one.

Corresponding to our data, we can call "sort" function to sort the unadjusted *p*-values as follows:

```
# call ''sort'' function
> sort(raw.pval)


(dose)1      (dose)0.6     (dose)0.2     (dose)0.05
0.007030475 0.015117195 0.057091821 0.500000000
```

This sorting of the unadjusted *p*-values just happened to result in reversing the order of the dosages from highest to lowest doses in these data, although this ordering of dose may not be the case for other dose-ranging trials. Then we proceed to compare the smallest *p*-value (i.e., 0.007 at dose 1) on whether it is smaller than 0.0125 (= 0.05/4, as there are 4 comparisons in the first step), which is true. We then proceed with a comparison to the next smallest *p*-value, which is 0.015 at dose 0.6, with 0.0167 (= 0.05/3, given 3 remaining comparisons), which is again true. We continue with this procedure to compare the next smallest *p*-value, which is 0.057 at dose 0.2 to 0.025 (= 0.05/2, at the remaining two comparisons), and can see that this is not true. We then stop the procedure and conclude that dose 1 and dose 0.6 are statistically significant. This Bonferroni–Holm procedure is implemented in R holm function and can be illustrated as follows:

```
# call ''holm'' function
> Holm.MCP = holm(raw.pval, alpha)


          Holm's (1979) step-down Procedure


Number of hyp.: 4
Number of rej.:  2
 rejected   pValues adjPValues
1      4 0.007030475 0.02812190
2      3 0.015117195 0.04535158


# print the output
> Holm.MCP

$adjPValues
(dose)0.05   (dose)0.2   (dose)0.6    (dose)1
0.50000000 0.11418364 0.04535158  0.02812190

$rejected
(dose)0.05 (dose)0.2  (dose)0.6   (dose)1
FALSE      FALSE     TRUE       TRUE
```

```
$criticalValues
[1] 0.01250000 0.01666667 0.02500000 0.05000000

$errorControl
An object of class ''ErrorControl''
Slot ''type'':
[1] ''FWER''

Slot ''alpha'':
[1] 0.05
```

Again notice that the R implementation is to expand the unadjusted p-values by the number of comparisons remaining and then to compare these adjusted p-values to the overall FWER at 0.05.

10.4.4 Hochberg Procedure

As one of the most popular multiple-comparison procedures in biopharmaceutical applications, the Hochberg procedure begins by ordering the unadjusted p-values from the largest to the smallest and then compares these p-values to a decreasing significance level given by $0.05/k$ at the kth step. Different from the Bonferroni-Holm procedure, the Hochberg procedure continues the testing until a statistical significance is reached; otherwise it will be concluded that none of these doses is statistically different from the placebo.

For our data, the unadjusted raw p-values have already been in the order from largest to the smallest corresponding to the doses at 0.05, 0.2, 0.6 and 1. The largest p-value of 0.5 at dose 0.05 is then compared with 0.05, which is larger than 0.05. We proceed to the second highest p-value at 0.057 at dose 0.2. We compare this p-value to 0.025 (= 0.05/2, at the second step) on whether it is larger, which is true again. We continue with the next smallest p-value, given by the third dosage. The smallest p-value at this dose is 0.015, which is used to compare with 0.0167 (0.05/3, at the third step). Since 0.015 is now smaller than 0.0167, we then stop and conclude a statistically significant difference between the dose at 0.6 and the placebo, which also implies automatically that there is a statistically significant difference between the higher dose at 1 and placebo. This Hochberg procedure is implemented in R hochberg function as follows:

```
# Call ''Hochberg'' function
> Hoch.MCP = hochberg(raw.pval, alpha)
```

```
          Hochberg's (1988) step-up procedure

Number of hyp.: 4
Number of rej.:  2
 rejected   pValues adjPValues
1        4 0.007030475 0.02812190
2        3 0.015117195 0.04535158
```

From the above output, we can see that there are two rejected tests from the four hypotheses, where the rejected hypotheses are at the third and fourth tests. We can then print the detailed results from this procedure as follows:

```
# print the detailed output
> Hoch.MCP

$adjPValues
(dose)0.05   (dose)0.2   (dose)0.6    (dose)1
0.50000000  0.11418364  0.04535158  0.02812190


$criticalValues
[1] 0.01250000 0.01666667 0.02500000 0.05000000

$rejected
(dose)0.05  (dose)0.2  (dose)0.6   (dose)1
   FALSE    FALSE      TRUE      TRUE

$errorControl
An object of class ''ErrorControl''
Slot ''type'':
[1] ''FWER''


Slot ''alpha'':
[1] 0.05
```

Again notice that the R implementation is to expand the unadjusted *p*-values by the number of steps. These adjusted *p*-values are then compared with the overall FWER at 0.05.

10.4.5 Gatekeeping Procedure

As discussed in Bauer and Budde (1994) and Bauer et al. (1998), this procedure is the so-called pre-determined step-down or the hierarchy procedure. Specifically, this procedure follows a pre-specified sequence where the comparison is to be conducted at the FWER of 0.05 level at each stage and continues as long as the

p-value is significant at the 0.05 level. The comparison is to be stopped whenever a *p*-value is above 0.05. In this sense, all comparisons are conducted at the level of 0.05, if the previous null hypothesis was rejected.

Because of this step-down procedure, it is very commonly used in dose-response analysis when a prior monotonic dose-response relationship can be assumed. For the postulated monotonic dose-response relationship, it is then logical to start with the highest dose by comparing the highest dose with the control and step-down on the doses against control.

In our monotonic example, we sort the unadjusted *p*-values from highest dose to lowest dose and compare these sorted *p*-values with 0.05 as the testing sequence. We can see that the *p*-value at the highest dose (i.e., dose = 1) is 0.007, which is smaller than 0.05, and we proceed to the next highest dose at dose = 0.6, where the *p*-value is 0.015, which is then less than 0.05. We again proceed to the next dose (0.2), where the *p*-value is 0.057, which is greater than 0.05. The procedure would stop here and conclude that the dose at 0.6 is statistically significant as is the dose at 1.0, the same conclusion reached as the previous procedures. In general, if a monotonic dose-response relationship cannot be assumed at the study design stage, the advice is not to use the gate-keeping procedure.

10.4.6 MCP Using p*-Values from Cochran-Mantel-Haenszel Test*

Cochran-Mantel-Haenszel (CMH) test is a commonly used test in biopharmaceutical industry. In dose-ranging study, the common practice is to perform a CHM analysis with only placebo and one dose, by excluding data from other dose treatment groups. Corresponding to these data, we would perform four CMH tests for dose combinations of (0, 0.05), (0, 0.2), (0, 0.6) and (0, 1). To limit the length of this chapter, we illustrate the CMH test for the last combination of (0, 1) and leave the other three combinations for interested readers.

To perform the CMH for the dose combination of (0,1), we first extract the corresponding data using following R code:

```
# Extract the data from ''biom''
> dd = biom[(biom$dose==0)| (biom$dose==1),]
```

Since we have 20 observations for each dose level, we would have 40 observations with three variables for this dose combination which can be seen as follows:

```
> dim(dd)
[1] 40  3
```

We can then create a contingence table using R function "table" as follows:

```
> # using contingency table, first create the table
> dat.tab =data.frame(table(dd$dose,dd$cresp))
# name the variables from the variables
> colnames(dat.tab) = c(''dose'',''better'',''cresp'')
# print the contingence table
> dat.tab

 dose better cresp
   0      0   14
   1      0    6
   0      1    6
   1      1   14
```

With this output, we can create a data frame to be used for R function "cmh_test". This can be done by using R function "xtabs" as follows:

```
# call ''xtabs'' to create data frame
> out.dat = xtabs(cresp~dose+better, dat.tab)
# print the data frame
> out.dat

   better
dose 0  1
  0 14  6
  1  6 14

# now call ''cmh_test'' to do the CMH test
> cmh_test(out.dat)

      Asymptotic Generalized Cochran-Mantel-Haenszel Test

data:  better by dose (0, 1)
chi-squared = 6.24, df = 1, p-value = 0.01249
```

It can be seen the two-sided *p*-value from CMH test for this dose combination is 0.01249, which is close to the *p*-value (i.e., 0.0141) from the logistic regression in Sect. 10.3.4. The associated one-side *p*-value is then 0.006245, which corresponds to the last value of 0.007030475 in Sect. 10.4.1.

Readers can follow the above steps for other dose combinations which would produce the one-sided raw *p*-values as follows:

```
> raw.pval
(dose)0.05  (dose)0.2  (dose)0.6  (dose)1
[1] 0.500000 0.057150  0.014315  0.006245
```

It can be seen that these raw p-values are quite close to the raw p-values at the end of Sect. 10.4.1 from the logistic regression in Sect. 10.3.4, especially for the first two dose combinations of (0, 0.05) and (0, 0.2).

With these raw p-values from CMH, we can reproduce the adjusted p-values in multiple comparison for procedures from Sect. 10.4.2 to Sect. 10.4.5. The detailed methodological explanations can be seen in Sects. 10.4.2–10.4.5 and we only re-produce the results here as comparison.

Corresponding to the procedure of Bonferroni adjustment at Sect. 10.4.2, the output is as follows:

```
> Bon.MCP = bonferroni(raw.pval, alpha, silent=FALSE)

          Bonferroni correction

Number of hyp.: 4
Number of rej.:  1
 rejected pValues adjPValues
       4 0.006245   0.02498

# print the detailed Bonferroni adjustment
> Bon.MCP
$adjPValues
[1] 1.00000 0.22860 0.05726 0.02498

$rejected
[1] FALSE FALSE FALSE TRUE
```

As seen from above, even the values are slightly different, the conclusion is the same that only dose at 1 is statistically significantly different from placebo.

Corresponding to the Bonferroni-Holm procedure in Sect. 10.4.3, the R output is as follows which gives the same conclusion:

```
# call the Bonferroni-Holm procedure
> Holm.MCP = holm(raw.pval, alpha)

          Holm's (1979) step-down Procedure

Number of hyp.: 4
Number of rej.:  2
 rejected pValues adjPValues
       4 0.006245  0.024980
       3 0.014315  0.042945
```

```
> Holm.MCP
   $adjPValues
   [1] 0.500000 0.114300 0.042945 0.024980

   $rejected
   [1] FALSE FALSE  TRUE  TRUE

   $criticalValues
   [1] 0.01250000 0.01666667 0.02500000 0.05000000
```

Corresponding to the Hochberg procedure in Sect. 10.4.4, the R output is as follows which gives the same conclusion:

```
   # call Hochberg procedure
   > Hoch.MCP = hochberg(raw.pval, alpha)

            Hochberg's (1988) step-up procedure

   Number of hyp.: 4
   Number of rej.:  2
    rejected pValues adjPValues
   1        4 0.006245  0.024980
   2        3 0.014315  0.042945
   > Hoch.MCP
   $adjPValues
   [1] 0.500000 0.114300 0.042945 0.024980

   $criticalValues
   [1] 0.01250000 0.01666667 0.02500000 0.05000000

   $rejected
   [1] FALSE FALSE  TRUE  TRUE
```

Again, corresponding to the gatekeeping procedure in Sect. 10.4.5, the R output is as follows which gives the same conclusion:

```
# GateKeeping by Dmitrienko et al 2006:
> GK.MCP =  raw.pval < alpha
> GK.MCP
[1] FALSE FALSE  TRUE  TRUE
```

10.5 Discussion

In this chapter, we illustrated both the modeling approach and the multiple-testing approach using the same dataset from Chap. 9 by dichotomizing the response data at the global median to illustrate the analysis and interpretation of a binary dose-response relationship. In modeling approach, we illustrated 4 methods to analyze dose-ranging studies and to establish a monotonic dose-response relationship. We further illustrated the estimation of effective doses within the dosing range that has a desirable probability to produce certain response using the fitted logistic regression model.

As illustrated when multiple doses are included in a dose-ranging trial, another question is to test the efficacy of each dose against the control (often a placebo) where the statistical analyses should be adjusted for multiple comparisons to control the family-wise Type 1 error rate. Many multiple testing procedures have been proposed in clinical trials which can be classified generally into two classes. The first class includes procedures that are developed specifically for continuous data, such as the Dunnett's method (1955) discussed in Chap. 9. The second class is for procedures that are 'distributional free' in the sense that their implementation does not depend on any particular distributional assumption, which is directly to use the unadjusted *p*-values from each test. For this, the procedures in the second class are readily applicable to categorical data, which is why we illustrated four procedures —Bonferroni, Bonferroni-Holm, Hochberg and the gate-keep procedure—in this chapter.

Throughout the chapter, the analysis and illustration of data were implemented in R software in a step-by-step fashion, so that interested readers can follow these steps to analyze their own data. For readers interested in SAS, please see Chuang-Stein and Li (2006).

As a final discussion, we would like make it clear that we do not promote routinely dichotomizing continuous data into binary data, which was done here simply for pedagogical purposes. Distinctions exist from strict binary responses (e.g., dead or alive) versus artificially created binary responses from continuous outcomes. Debating on the pros and cons of dichotomization, Lewis (2004) stated that clinical researchers often simplify and dichotomize continuous data to make sense of unfamiliar clinical measurements and treatment effects of uncertain implication with a defendable threshold value based on continuous measurements. Well-known examples in dichotomization in real life can be understood to dichotomize the diastolic blood pressure at 80 mmHg as a borderline to determine normal/high blood pressure and a total cholesterol level at 200 mg/dL as a borderline to determine the desirable/undesirable cholesterol level.

The cons for this dichotomization is easily seen to be too crude in real practice, but the pros can also be seen as this simplicity can really help human minds to make decisions easily since human decisions are often indeed binary in nature. In an extensive examination, MacCallum et al. (2002) concluded and warned that dichotomization has substantial negative consequences in losing of information

about individual differences and loss of effect size and power in statistical analysis among other negative consequences. Further discussions can be seen from Snapinn and Jiang (2007) and Uryniak et al. (2012). We encourage the readers to study these papers and be cautious about dichotomizing the data for statistical analysis. In this chapter, we dichotomize the `biom' data just for the purpose of continuation from the previous chapter and illustrate the methods for analyzing the dose-ranging clinical trials.

References

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). New York: Wiley.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics, 11*(3), 375–386.

Bauer, P., & Budde, M. (1994). Multiple testing for detecting efficient dose steps. *Biometrical Journal, 36,* 3–15.

Bauer, P., Rohmel, J., Maurer, W., & Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine, 17,* 2133–2146.

Bretz, F., Pinheiro, J. C., & Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, *61*(3), 738–748.

Chen, D. G., & Peace, K. E. (2011). *Clinical trial data analysis using R*. Boca Raton, FL: Chapman and Hall/CRC Biostatistics Series.

Chen, D. G., Peace, K. E., & Zhang, P. (2017). *Clinical trial data analysis using R and SAS* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC Biostatistics Series.

Chuang-Stein, C., & Li, Z. (2006). Analysis of dose response relationship based on categorical outcomes. In Ting, N. (Ed.), *Dose finding in drug development* (Chap. 13, pp. 200–219). Berlin: Springer.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics, 10*(4), 417–451.

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

Lewis, J. A. (2004). In defence of the dichotomy. *Pharmaceutical Statistics*, *3*(2), 77–79.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society: Series B, 42,* 109–142.

Newcombe, R. G. (1998a). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17,* 857–872.

Newcombe, R. G. (1998b). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine, 17,* 873–890.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5 50*(302), 157–175.

Snapinn, S. M., & Jiang, Q. (2007). Responder analysis and the assessment of a clinically relevant treatment effect. *Trials, 8*(31), 1–6.

Uryniak, T., Chan, I. S. F., Fedorov, V. V., Jiang, Q., Oppenheimer, L., Snapinn, S. W., et al. (2012). Responder analysis-A PhRMA position paper. *Statistics in Biopharmaceutical Research, 3*(3), 476–487.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of American Statistical Association, 22,* 209–212.

Chapter 11
Bayesian Approach

11.1 Introduction

Generally there are two approaches to statistical inference, namely the frequentist and Bayesian approaches. The frequentist approach considers probability as a limiting long-run frequency, whereas the Bayesian approach regards probability as a measure of the degree of belief about the value of an unknown parameter. As such, Bayesian statistics expresses uncertainty about unknown parameters probabilistically. Bayesian inference is a method of updating the probability for a hypothesis as more evidence or information becomes available. In this chapter, we will illustrate the Bayesian concept, update and inference by using a simple example. We will also discuss its application to Phase II clinical trials by continuing the dose-finding example used in Bretz et al. (2005) from Chap. 9.

11.1.1 An Example on Bayesian Concept

Let's consider a simple open-label clinical trial that involves only one treatment Q to treat headache (referred as the Q example hereafter). There was some anecdote evidence that treatment Q is effective. So we may conduct a trial to test the hypothesis of Q's response rate p using a number of subjects. Furthermore, let's recruit subjects one at a time so each subject's response to Q is available before the next subject is recruited. Before the trial started, one may have some subjective opinion about Q. For example, Q can be perceived just like a popular over-the-counter (OTC) medicine for headache with a strong response rate, or Q is similar to an herb remedy that only works for certain people. Suppose we have three researchers, Captain is a strong believer in Q and perceives the actual $p = 0.9$, Spock is neutral and perceives that $p = 0.5$, while Yeoman is strongly suspicious and perceives the actual $p = 0.1$.

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_11

Now if the first outcome is positive, Captain might feel fine about his perception and Yeoman might think that it just happened to be one of those rare positive cases. If the second outcome is also positive, Captain might become more confident, while Spock might still feel fine for his neutral perception but Yeoman might become less confident about her perception. If the third outcome is negative, Captain might lose a little confidence while Spock and Yeoman might gain some confidence. On the other hand, if the third outcome is still positive, Captain might gain more confidence while Spock and Yeoman might lose some confidence. This can continue until the trial is over. As one can imagine, when information accumulates, the confidence of subjective belief may change. Initially the belief is mostly subjective; however, as information accumulates, the belief becomes less subjective because more information through empirical data has been incorporated into the belief. How can the belief be quantified and updated? How does this idea work as a statistical method? The answer is Bayesian statistics and the Bayesian method.

11.1.2 A Brief History

The term *Bayesian* came from Bayes' theorem developed by Thomas Bayes (1702–1761). The theorem is rather simple yet elegant and it is based on the principle of conditional probability. Pierre-Simon Laplace (1749–1827) later introduced a general version of the theorem and used it to approach real problems. Bayesian statistics actually predates frequentist statistics, but only became more popular for the last few decades during which computing speed and power increased exponentially.

In the 1980s, the discovery of Markov Chain Monte Carlo methods (described later in this chapter) relieved many of the computational burdens and enabled Bayesian methods usage for nonstandard and complex applications.

11.1.3 Bayes Theorem

Bayes Theorem is based on the concept of conditional probability as described below.

Given two events A and B with $P(B) > 0$, the conditional probability of A given B is defined as the ratio of the joint probability of A and B, and the probability of B:

$$P(A|B) = P(A \cap B)/P(B)$$

Essentially this restricts the sample space to B and it describes the conditional probability of A given B as the ratio of probability of both events and the probability of event B.

Similarly with $P(A) > 0$, the conditional probability of B given A is defined as the ratio of the joint probability of A and B, and the probability of B:

$$P(B|A) = P(A \cap B)/P(A)$$

From the above two conditional probabilities, it's clear that

$$P(B|A) = P(B)P(A|B)/P(A)$$

This is exactly the equation for Bayes theorem in the sense that, if event A occurs before event B, then the posterior probability (probability of event B after event A occurred) is a consequence of a prior probability (probability of event B before event A occurs) and a conditional probability for event A given event B which is yet to occur, normalized by the (total) probability of event A. The less intuitive and intriguing concept is the probability conditional on a future unknown event, which is useful when applied to the framework of hypothesis testing using posterior probability, conditioned on yet to be observed future data, as described below.

11.1.4 Bayesian Hypothesis Testing Framework

Let's first very briefly review the frequentist hypothesis testing. It starts with a null hypothesis H_0 in terms of parameter values, for instance, H_0: $\theta = \theta_o$, then collects data D, and uses the sampling distribution of the data given the parameter value (typically the null value of no effect), $P(D|H_0)$, to evaluate how likely H_0 is true in order to decide whether or not to reject the H_0.

The Bayesian approach starts with the prior distribution of the parameter, $P(\theta)$. After data are collected, Bayes theorem is applied to derive the posterior distribution of parameter, $P(\theta|D)$, and then to make statistical inferences about θ. The prior distribution can be updated as more data are cumulated, and a new posterior distribution can be derived for updating statistical inferences. In this manner Bayesian approach follows a general scientific principle—that is, use of the cumulative information or knowledge to make inferences.

To derive the posterior probability $P(\theta|D)$, again we apply the Bayes' theorem:

$$P(\theta|D) = P(\theta)P(D|\theta)/P(D)$$

where

- θ stands for any hypothesis whose probability may be affected by data;
- D corresponds to data yet to be observed and therefore have not been used in contributing to the prior probability;
- $P(\theta)$, the prior probability, is the probability of θ before D is observed; it's the prior belief about how likely different hypotheses are;

- $P(\theta|D)$, the posterior probability, is the probability of θ after D is observed; it's the probability of various hypotheses about θ given the observed data;
- $P(D|\theta)$, the probability of observing D *given* θ, is also known as the likelihood; it indicates the compatibility of the data with the given hypothesis;
- $P(D)$ is the marginal likelihood, the integration of $P(D|\theta)$ over θ; this factor is the same for all possible hypotheses being considered and does not enter into determining the relative probabilities of different hypotheses.

Note that the posterior probability of a hypothesis is determined by a combination of the prior belief of the likeliness of a hypothesis and the compatibility of the observed data with the hypothesis (likelihood). This also means that the posterior is proportional to prior times likelihood. The critical point about Bayesian method is that it provides a way of combining new data with prior beliefs, through the application of Bayes' theorem. Furthermore, Bayes' theorem can be applied iteratively: after observing some data, the resulting posterior probability can then be treated as a prior probability, and a new posterior probability computed from new data. This allows for Bayesian principles to be applied to various kinds of data, whether viewed all at once or over time.

11.2 Bayesian Updating

11.2.1 Example Continued for Bayesian Updating

Now let's consider the Q example, this time we will see how Bayes theorem is used to update Bayesian (posterior) probabilities. For illustrative purposes, let's assume that there are only three possible response rates of treatment Q and they are 0.1, 0.5, and 0.9. Essentially we are assuming that there are three hypotheses about *p,* and they are Captain's $p = 0.9$, Spock's $p = 0.5$, and Yeoman's $p = 0.1$. Let's also define a subject outcome as *Success* if headache disappears within one and half hours after treatment or *Failure* otherwise (i.e., headache persists one and half hours after treatment).

For the prior probability distribution, it is a probability function of the response rate expressing the degree of belief about it, prior to observing data. In this case, to start with, we may consider that all three hypothetical probabilities of a successful outcome are equally likely. In other words, Captain, Spock and Yeoman start with an equally likely prior for p, so that the initial $P(p)$ is the same for the three possible values or hypotheses of p. That is,

$$P(p) = \frac{1}{3} \, for \, p = 0.1, 0.5, 0.9.$$

For the likelihood function $P(D|\theta = p)$, the probability of observing D *given* P, follows a Bernoulli distribution:

$$P(D|p) = p \text{ when the event is a Success},$$
$$P(D|p) = 1 - p \text{ when the event is a Failure}$$

Let's now assume that the first 5 subject outcomes are as follows: *Success, Failure, Success, Success,* and *Failure*. The Bayes theorem $P(p|D) = P(p)P(D|p)/P(D)$ can be used to update the posterior probability after each outcome is observed.

The first outcome is a *Success*. Hence:

For Captain's posterior probability, $P(p|D)_1 = P(p)P(D|p)/P(D)$ where $P(p) = 1/3$ and $P(D|p) = 0.9$, so $P(p|D)_1 = (1/3) * 0.9/P(D)$. Similarly for Spock's posterior probability, $P(p|D)_1 = (1/3) * 0.5/P(D)$, and for Yeoman's posterior probability, $P(p|D)_1 = (1/3) * 0.1/P(D)$. Here $P(D)$ is a normalizing factor such that the sum of all $P(p|D)_1$'s equals 1, which is the total probability of any probability distribution, including any posterior probability distribution. Equivalently, $P(D)$ is the sum of $P(p)P(D|p)$ over all possible p's, that is, $(1/3) * 0.9 + (1/3) * 0.5 + (1/3) * 0.1 = 0.5$. Now the posterior probability distribution among the three hypotheses can be obtained as:

$$\textit{Captain}: \ P(p = 0.9|D)_1 = \frac{9}{15}, \quad \textit{Spock}: \ P(p = 0.5|D)_1 = \frac{5}{15},$$
$$\textit{Yeoman}: \ P(p = 0.1|D)_1 = \frac{1}{15}.$$

Notice the posterior probability differences after just a single event, and now this distribution serves as a new prior distribution for the next event.

The second outcome is a *Failure*. Hence:

For Captain's posterior probability, $P(p|D)_2 = P(p|D)_1 P(D|p)/P(D)$ where $P(p|D)_1 = 9/15$ and $P(D|p) = 1 - 0.9 = 0.1$, so $P(p|D)_2 = (9/15) * 0.1/P(D) = (3/50)/P(D)$. Similarly for Spock's, $P(p|D)_2 = (5/15) * (1 - 0.5)/P(D) = (1/6)/P(D)$, and for Yeoman's, $P(p|D)_2 = (1/15) * (1 - 0.1)/P(D) = (3/50)/P(D)$. Here $P(D)$ again is the sum of $P(p|D)_1 P(D|p)$ over all $P(p|D)_1$'s and that is (3/50) + (1/6) + (3/50) = 43/150. Now the posterior probability distribution among the three hypotheses can be obtained as follows:

$$\textit{Captain}: \ P(p = 0.9|D)_2 = \frac{9}{43}, \quad \textit{Spock}: \ P(p = 0.5|D)_2 = \frac{25}{43},$$
$$\textit{Yeoman}: \ P(p = 0.1|D)_2 = \frac{9}{43}.$$

Notice the swinging effect after the second outcome being a *Failure*; Captain and Yeoman now have the same posterior probability about their individual hypothesis, while Spock has gained more posterior probability than both of them. This situation occurs because the current data supports 50–50 ($p = 0.5$) hypothesis more than the other two hypotheses. Now this distribution serves as the new prior distribution for the next event.

The third outcome is a *Success*. Hence:

For Captain's posterior probability, $P(p|D)_3 = (9/43) * (0.9)/P(D) = (81/430)/P(D)$; for Spock's, $P(p|D)_3 = (25/43) * (0.5)/P(D) = (125/430)/P(D)$; and for Yeoman's, $P(p|D)_3 = (9/43) * (0.1)/P(D) = (9/430)/P(D)$.

Here $P(D)$ is the sum of $P(p|D)_2 P(D|p)$ over all $P(p|D)_2$'s, and that is (81/430) + (125/430) + (9/430) = 215/430. Now the posterior probability distribution among the three hypotheses can be obtained as follows:

$$Captain: \ P(p = 0.9|D)_3 = \frac{81}{215}, \quad Spock: \ P(p = 0.5|D)_3 = \frac{125}{215},$$
$$Yoeman: \ P(p = 0.1|D)_3 = \frac{9}{215}.$$

Again it can be seen how the posterior probabilities change as more data are obtained; it's like a tug of war among all possible hypotheses. Now this distribution serves as the new prior distribution for the next event and the iteration continues until no more data are retrieved. Table 11.1 summarizes the posterior probabilities after each outcome is observed.

From this limited number of outcomes and restricted hypotheses, the posterior probability distribution after 5 outcomes seems to support that this treatment Q may have a 50–50% chance of stopping a headache within one and half hours because it has the largest posterior probability of 0.79. **This can also be generalized to say that the Bayesian decision is simply for the hypothesis with the largest posterior probability**.

Table 11.1 Posterior probabilities for an outcome sequence: *Success, Failure, Success, Success,* and *Failure*

Outcome	Posterior probabilities for Captain's hypothesis ($p = 0.9$)	Posterior probabilities for Spock's hypothesis ($p = 0.5$)	Posterior probabilities for Yeoman's hypothesis ($p = 0.1$)
0. Before any event	1/3 (0.33)	1/3 (0.33)	1/3 (0.33)
1. *Success*	9/15 (0.60)	5/15 (0.33)	1/15 (0.07)
2. *Failure*	9/43 (0.21)	25/43 (0.58)	9/43 (0.21)
3. *Success*	81/215 (0.38)	125/215 (0.58)	9/215 (0.04)
4. *Success*	729/1363 (0.53)	625/1363 (0.46)	9/1363 (0.01)
5. *Failure*	729/3935 (0.19)	3125/3935 (0.79)	81/3935 (0.02)

Table 11.2 Posterior probabilities for an outcome sequence of 5 *Successes*

Outcome	Posterior probabilities for Captain's hypothesis (p = 0.9)	Posterior probabilities for Spock's hypothesis (p = 0.5)	Posterior probabilities for Yeoman's hypothesis (p = 0.1)
0. Before any event	1/3 (0.33)	1/3 (0.33)	1/3 (0.33)
1. *Success*	9/15 (0.60)	5/15 (0.33)	1/15 (0.07)
2. *Success*	81/107 (0.76)	25/107 (0.23)	1/107 (0.01)
3. *Success*	729/855 (0.85)	125/855 (0.15)	1/855 (<0.01)
4. *Success*	6561/7187 (0.91)	625/7187 (0.09)	1/7187 (<0.01)
5. *Success*	59049/62175 (0.95)	3125/62175 (0.05)	1/62175 (<0.01)

To further provide a flavor of the Bayesian method, Table 11.2 illustrates posterior probabilities when there are five consecutive successes. From here one can see how quickly Captain's belief increases while Spock's and Yeoman's belief fades away.

It is important to note that it is possible that the Bayesian update can be done after a batch of outcomes become available. Yet the posterior probabilities will not change due to the frequency of updates. This situation can occur when the prior and the data are exactly the same, regardless how frequent the posterior probabilities get calculated. For the example above, we may do the first update when three outcomes are available and then the second update when the remaining two outcomes are available. In this case, we just need to consider the likelihood function as being a Binomial distribution (note that Bernoulli is a special case of Binomial when there is only one event), because we will be updating after a few events; that is, we will be counting how many *Successes* and *Failures* among the first three and on the last two outcomes. So now the likelihood function $P(D|p)$, the probability of observing D *given* P, for the first three outcomes, follows a Binomial distribution:

$$P(D|p) = \binom{3}{k} p^k (1-p)^{3-k}$$

for k = 0, 1, 2, 3 *Successes*, where $\binom{3}{k} = \frac{3!}{k!(3-k)!}$ and P is the probability of a *Success*.

For the first update after 2 *Successes* and 1 *Failure,* the posterior probability for Captain is $P(p|D)_1 = P(p)P(D|p)/P(D)$ where $P(p) = 1/3$ and $P(D|p) = \binom{3}{2} p^2(1-p)^{3-2} = 3(0.9)(0.9)(0.1)$, so $P(p|D)_1 = (243/1000)/P(D)$. Similarly for Spock, $P(D|p) = 3(0.5)(0.5)(0.5)$, so $P(p|D)_1 = (375/1000)/P(D)$; and for Yeoman, $P(D|p) = 3(0.1)(0.1)(0.9)$, so $P(p|D)_1 = (27/1000)/P(D)$. Since $P(D)$ is just a normalizing factor so the posterior probabilities for Captain, Spock, and Yeoman are proportional to 243/1000, 375/1000, and 27/1000 respectively and they

need to sum up to 1. As a result, the posterior probabilities for Captain, Spock, and Yeoman can be obtained as 243/(243 + 375 + 27) = 243/645 = 81/215, 375/645 = 125/215, and 27/645 = 9/215, respectively. Note that these are identical to the third update in Table 11.1.

It will be a similar and simple exercise to use these updated Bayesian probabilities as the new prior and update the Bayesian probabilities after the fourth and fifth outcomes. The posterior probabilities will be identical to the fifth update shown previously.

All of the above is just a simple example to illustrate the Bayesian concept and computation. Typically the hypotheses, likelihood, prior distribution, and event outcomes can be much more complicated. The Bayesian posterior probability derivation or computation can be intensive and requires heavy computation power. This computational burden was one of the major obstacles for the Bayesian methods until the last few decades. Nowadays with the tremendous advances in computing technology, the Bayesian method can be implemented relatively quickly and easily, as described in subsequent sections.

11.3 Bayesian Inference

Similar to the frequentist inference, Bayesian inference also includes hypothesis testing and interval estimation. For Bayesian hypothesis testing, the above example illustrates an oversimplified case. Nonetheless the principle is to use posterior distribution to calculate the probability that a particular hypothesis is true, given the observed data. Bayesian interval estimates are based on the posterior distribution and are often called *credible intervals*.

If the posterior probability that an endpoint (e.g., a treatment response rate) lies in an interval is 0.95, then this interval is called a 95% *credible interval*. For the Q example, since we assume that there are only three possible response rates (0.1, 0.5, and 0.9) with three discrete posterior probabilities, the credible interval for the response rate does not have much meaning. However in reality, the response rate can be any value between 0 and 1, therefore a wider range posterior distribution for the response rate can be obtained and so can *credible intervals*. To further elaborate using the Q example, suppose the prior distribution of response rate follows the Uniform [0, 1] distribution, that is, any value between 0 and 1 is equally likely to be the true response rate. This is one kind of "non-informative" prior for not giving any preference to any possible response rates. The likelihood of n responses follows the binomial distribution. In this case, the posterior probability distribution follows a Beta distribution as derived below.

Consider the number of positive responses $k \sim$ Binomial (n, p) where n is the total number of subjects participating (sample size) and p is the response rate parameter. Then as given in the previous section the likelihood function is

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

for k = 0, 1, 2, …, n, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

The prior of p, $P(p)$ follows Uniform (0, 1), that is, $P(p) = 1 \, for \, 0 < p < 1$ and $P(p) = 0$ otherwise. This prior is also a Beta (a = 1, b = 1) distribution because the definition of Beta distribution is given by

Beta (a, b) $= \frac{\Gamma(\mathrm{a})\Gamma(\mathrm{b})}{\Gamma(\mathrm{a}+\mathrm{b})} p^{\mathrm{a}-1}(1-p)^{\mathrm{b}-1}$, where $\Gamma(\mathrm{c}) = (c-1)!$ for any positive integer c and 0! = 1. Furthermore, this prior is also a conjugate prior. A conjugate prior for the likelihood function is a prior that produces a posterior in the same distribution family. In this case, this Beta(1, 1) prior $P(p)$ for the binomial likelihood function $P(k|n,p)$ produces the posterior probabilities

$$P(p|k) \propto P(k|n,p) \times P(p) \propto p^k (1-p)^{n-k} \times 1 = p^{(k+1)-1}(1-p)^{(n-k+1)-1},$$

which follows a beta distribution with parameters k + 1 and $n - k$ + 1.

Therefore Bayesian inferences for p can be obtained using this posterior beta distribution.

For example, if n = 100 and k = 60, which means there were 100 subjects took the Q treatment and 60 of them had a positive response (similar to 60% response rate as for the Q example), then the posterior probability distribution follows Beta (60 + 1, 100 − 60 + 1) = Beta(61, 41).

It can be shown that the mean of Beta(a, b) is $a/(a + b)$, so this posterior mean for p is 61/102 = 0.598. It is noted that while one can also choose the posterior median or mode as the Bayesian point estimator, here we choose the posterior mean.

Now let's review the definition of the credible interval. It is an interval where an outcome of an endpoint lies in an interval with posterior probability of $1 - \alpha$. Therefore there can be different choices of credible intervals that satisfy this definition. Two choices are commonly used: (1) equal-tailed (1/2 α probability on each side) and (2) Highest Probability Density (HPD) interval under the posterior distribution. For symmetric unimodal posterior distribution, these two credible intervals are identical. Otherwise the HPD credible interval is often used for Bayesian approach, because it is the narrowest interval that covers the $1 - \alpha$ credible level and all values within it are more probable than other values outside of it. As the actual calculation of the HPD interval typically involves iterations (in order to determine the narrowest interval that covers $1 - \alpha$ probability) and requires a computing algorithm, it is easier to use computer software such as SAS Proc MCMC to obtain it.

Based on the Beta(61, 41), a 95% HPD for p can be obtained as (0.503, 0.692) (see Proc MCMC code in the next section), while the 95% equal-tailed interval is

(0.502, 0.691). The latter can be easily obtained by using the Excel Beta inverse function [(BETAINV (0.025, 61, 41), BETAINV (0.975, 61, 41)]. Also included here as a reference is the frequentist 95% Wald confidence Interval using a normal approximation

$$(\hat{p} - 1.96\sqrt{(\hat{p}(1-\hat{p})/n)}, \hat{p} + 1.96\sqrt{(\hat{p}(1-\hat{p})/n)}).$$

Given $\widehat{p} = 60/100$, this interval is (0.504, 0.696).

It is a simple exercise to obtain a posterior mean and 95% credible interval for the posterior probabilities of Captain, Spock and Yeoman, assuming that they agreed to use the Uniform (0, 1) as the prior for p and observed the same five responses with 3 successes and 2 failures. In such a case, in principle there can be infinite many hypotheses as p is assumed to take any value between 0 and 1;, however, here the context is just for the three competing hypotheses of Captain, Spock and Yeoman. The posterior distribution of p is then Beta(4, 3), with the posterior mean 4/7 = 0.571. The 95% equal-tailed and HPD credible intervals are (0.223, 0.882) and (0.243, 0.900), respectively. Although the two scenarios have the same 60% response rate, as expected, the one based on a much larger sample provides more precise inferences about p. One can also see that these intervals support Spock's $p = 0.5$ hypothesis, among the three competing hypotheses.

It is worth noting that, although the term "non-informative" is used for the Uniform (0, 1) prior, it does not mean no information at all is introduced. Non-informative prior is not unique and there are different criteria for choosing non-informative priors. For example, the Beta(0.5, 0.5) is actually less informative than the Uniform (0, 1) prior, as pointed out in Zhu and Lu (2004). It is less informative because the associated Bayesian posterior estimator is more similar to the Maximum Likelihood Estimator (MLE) of p. This and many other interesting Bayesian topics are beyond the scope of this chapter. Interested readers can refer to Berger (1985) and Kass and Wasserman (1996) for additional information.

11.4 Markov Chain Monte Carlo (MCMC) Method

MCMC is a sampling method for simulating data from the distributions of random quantities. It consists of two parts: the Markov Chain part and the Monte Carlo simulation part.

The Monte Carlo part, in the context of the MCMC method, refers to random simulations or draws from a probability distribution. It is well understood that by sufficient number of draws, the probability distribution can be approximated as close as the desired precision. Interestingly, the use of the term Monte Carlo was originated from classified nuclear weapons projects at the Los Alamos National Lab by Stanislaw Ulam and John von Neumann in the late 1940s. The project required a secret code name and a colleague of them, Nicholas Metropolis, suggested using the

name Monte Carlo, which refers to the Monte Carlo Casino in Monaco where Ulam's uncle would borrow money from relatives to gamble (Metropolis 1987).

A Markov Chain, in the context of the MCMC method, can be viewed as a sequence of random moves M1, M2, M3 and so on in the sample space with the probability of moving to next state depending only on the present state and not on the previous states. Hence the conditional probability distribution for the system at the next step and at all future steps depends only on the current state of the system and not additionally on the state of the system at previous steps. In theory, the chain needs to converge to a steady state, namely the target probability distribution, for it to be statistically useful. However, in practice, the convergence is often difficult to be definitively confirmed; therefore, satisfying reasonable convergence diagnostics criteria is sufficient for the MCMC results to be used.

Putting these two parts together, the MCMC method is to construct a sound Markov chain with a prescribed stationary probability distribution that can be parameterized, and then use Monte Carlo simulations to approximate this stationary probability distribution. Statistical inferences can then be performed by using the approximated probability distribution.

There are ways to construct a sound Markov chain, such as Gibbs sampler or Metropolis-Hastings (MH) algorithm. The MH algorithm can be very general by using a known and easier distribution (e.g., normal distribution) for simulation, and it also avoids the integration of the posterior distribution, which makes it attractive for Bayesian modeling. A general five-step MCMC algorithm is as follows:

1. Start at current place M_i
2. Propose moving to a new place M_{i+1}
3. Accept the new place based on the place's adherence to the data and prior distributions, typically by a probability criterion
4. If acceptable, move to the new place and return to Step 1
5. After a number of iterations, return the samples.

Essentially this algorithm leads samples to move towards the regions where the posterior distributions exist meanwhile samples are also collect concurrently. When the posterior distribution is reached (converged), the new samples drawn are likely to all come from the posterior distribution.

In summary, MCMC method is popular for Bayesian approach because in many analyses, the posterior distribution is too complex to write down and therefore traditional numerical integration techniques cannot be carried out. Recall the computations of our simple Q example given previously. Imagine how much more tedious and complex the posterior distribution would be when the prior and likelihood are complex. The MCMC sampler draws samples from other known distributions to create a Markov chain of values. Eventually, as the chain converges, the values sampled begin to resemble draws from the posterior distribution. The draws from the Markov chain can then be used to approximate the posterior distribution and make Bayesian inferences. Current statistics software such as SAS PROC MCMC has further made this method easy to implement. Below is a

sample code for generating the 95% credible intervals for the Beta(61, 41) posterior distribution, the example used in the previous section.

```
Title "Bayesian Credible Intervals for Binomial p by PROC
MCMC";
Data Qtrt; /*input dataset to specify 60 successes out of 100
events*/

      n=100; k=60;
      run;

Proc MCMC data=Qtrt;
seed=111
outpost=PostSample /*output dataset for posterior samples*/
nbi=10000 /*burn-in samples, to be discarded, optional for
this example*/
nmc=100000 /*number of iterations*/
statistics=all diagnostics=all plots=all; /*output all
relevant info*/
parms p 0.5; /*gives p an initial value*/
prior p ~ beta(1,1); /*uniform (0 1) prior of p*/
model k ~ binomial(n,p); /*likelihood function for k*/
run;
```

The primary output follows:

Bayesian credible intervals for binomial *p* by PROC MCMC			
Parameter	Alpha	Equal-tail interval	HPD interval
p	0.050	0.5016–0.6907	0.5026–0.6915

Interested readers are recommended to refer to Chen and Peace (2011), Gamerman and Lopes (2006), and Brooks et al. (2011) for additional information.

11.5 Bayesian Methods for Phase II Clinical Trials

Bayesian methods depend on prior information to help leverage what has been learned before in order to gain efficiency (by reducing uncertainties) in treatment inference and effect estimation, using less resources (smaller sample sizes). Although Bayesian decisions depend heavily on utility functions (which typically involve costs and benefits) in which the criteria for decisions are based upon, our focus is on Phase II clinical trials where objectives are more clear (e.g., dose finding) without more complicated utility issues. We will not discuss the utility function aspect of Bayesian methods.

Design and analysis methods for phase II dose-ranging clinical trials include multiple comparisons among doses, modeling of a dose-response relationship, and a hybrid approach such as MCPMod (Bretz et al. 2005). These approaches serve well when the data are positive and resources (e.g., sample size) are adequate. Given resource constraints but with efficient computer power, Bayesian dose-response modeling method offers an alternative. The dose-response modeling links multiple doses and shares information among doses, the Bayesian incorporates prior information from previous studies or experts' opinions; therefore, it may lead to more powerful comparisons and more precise estimates. The information-rich era and an increasing trend of clinical data sharing by pharmaceutical companies such as the SHARE program initiated by GlaxoSmithKline in 2013 (https://www.clinicalstudydatarequest.com), make it easier to acquire relevant prior information for clinical development that reaches Phase II. Therefore informative Bayesian priors—based on the best available evidence—can be obtained to help design and analyze Phase II dose-ranging clinical trials.

The challenge when leveraging historical data to form a prior for a new study is to determine whether the existing data are sufficiently similar to the setting of the new trial. For a Bayesian modeling Phase II clinical trial, prior distributions for the dose response model parameters are needed. One can always use non-informative priors to implement the Bayesian approach; nevertheless, informative priors may take advantage of the Bayesian methods, especially if one can find and use the most relevant historical data to estimate the dose-response curve shape and also to estimate the within- and between-study variability.

There are generally two ways of obtaining informative prior distributions. One way is to rely on specific historical data, which is also known as data-driven method. The other way is to use prior elicitation which is an exercise that combines historical data, literature and expert opinion to form a prior distribution. There are packages such as SHELF (The Sheffield Elicitation Framework), that do prior elicitation (Oakley and O'Hagan 2010) They can be used to facilitate discussion among team members and experts to obtain a consensus prior distribution, however, they do require training, meeting preparation, and time to conduct the elicitation. The prior elicitation methods will not be covered in this chapter.

11.6 Example

We continue the Chap. 9 example for the "biom" data in the R packages, to illustrate the Bayesian modeling approach. We adopt the Emax modeling for the data and introduce Bayesian priors for the Emax parameters. We first choose one set of non-informative priors to illustrate how the Bayesian modeling approach is implemented, and then randomly select a subset of "biom" data to construct a set of "informative" priors to explore differences in dose-response modeling and dose selection.

Let us first refresh the frequentist Emax modeling results for this example from Sect. 9.5:

E0 mean (SE), 95% C. I.	Emax mean (SE), 95% C. I.	ED50 mean (SE), 95% C. I.
0.322 (0.152), (0.024, 0.620)	0.746 (0.236), (0.283, 1.209)	0.142 (0.180), (0.0, 0.500)[a]

[a]Adjusted from (−0.216, 0.500) due to ED50 > 0

11.6.1 Using Non-informative Priors

There are three parameters, E_0, E_{max}, ED_{50} that need priors. For E_0 and E_{max}, in principle they can be both positive and negative, but for ED_{50}, it is a dose in the dose range scaled between 0 and 1. Without any other information, we can use the following non-informative independent priors as shown in Table 11.3.

Note in Table 11.3 that we choose a normal distribution for both E_0 and $E_{\max}$ with mean 0 and a large standard deviation of 10 so that the distribution is flat relative to the response range. This essentially provides non-informative priors for these two parameters. For the ED_{50}, we choose a beta-distribution in the interval of 0 to 1 with the two shape parameters to be 0.5 as a non-informative prior. As mentioned earlier in this chapter, this prior is less informative than the Uniform (0, 1) = Beta(1, 1) in terms of the Bayesian posterior estimator being less different from the MLE estimator of p (Zhu and Lu 2004).

Given these priors, we perform the Bayesian modeling to the data with R package "DoseFinding" using the following R code chunk:

```
### We first produce the mean estimates at each dose with "lm"
>mod4Mean= lm(resp~factor(dose)-1, data=biom)
# Extract the estimated mean and covariance matrix
>estMean = coef(mod4Mean)  # the estimated means
        S = vcov(mod4Mean) # the covariance matrix


### Now Bayesian analysis with non-informative prior
>prior1 <- list(norm = c(0, 10), norm = c(0,10),
beta=c(0,1,0.5,0.5))
# Fit the Emax model using Bayesian with 10,000 MCMC
>b1Mod <- bFitMod(dose, estMean, S, model = "emax",
                 nSim = 10000, prior = prior1)
```

Table 11.3 Non-informative priors for the Emax model

Parameter	Non-informative prior distribution
E_0	Normal with mean 0 and standard deviation 10
E_{max}	Normal with mean 0 and standard deviation 10
ED_{50}	Beta-distribution in (0, 1) with both shape parameters to be 0.5

```
# Print the summary information
>print(b1Mod)
```

As can be seen in the above code, we made a `list` of priors and named it `prior1`. For Bayesian modeling, we can use the R function `bFitMod` (in short for "Fit a dose-response model using Bayesian or bootstrap methods") in `DoseFinding` package with options: `model = "emax", nSim = 10000, prior = prior1` for 10,000 MCMC samples. For the 10,000 MCMC samples, the output summary information is as follows:

Summary of posterior draws							
	mean	sdev	2.5%	25%	50%	75%	97.5%
e0	0.386	0.132	0.1097	0.300	0.389	0.476	0.629
eMax	0.931	0.338	0.3423	0.698	0.900	1.136	1.680
ed50	0.513	0.326	0.0327	0.202	0.493	0.835	0.997

This summary information of the posterior distributions for these three parameters are reported as the mean and standard deviation, as well as the sample quantiles of 2.5, 25, 50, 75 and 97.5%. This Bayesian model can be graphically illustrated in Fig. 11.1. Generally this figure is similar to Fig. 9.3 in Chap. 9.

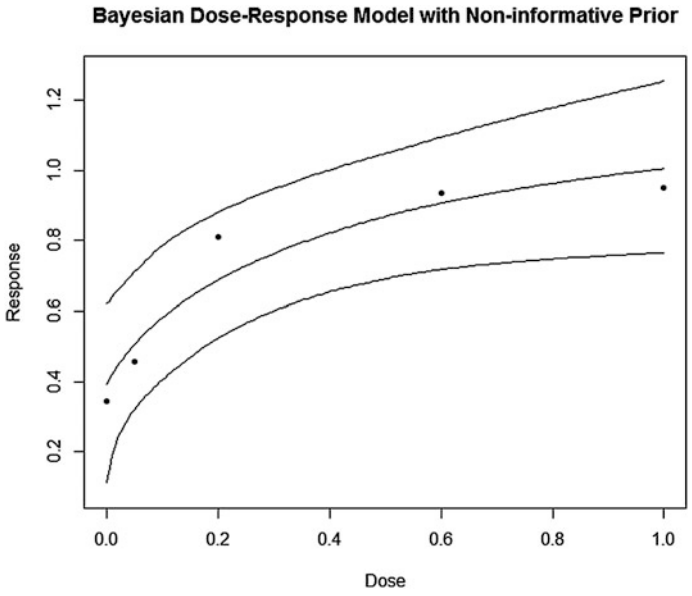

Fig. 11.1 Bayesian dose-response model with non-informative priors

11.6.2 Using Informative Priors

Because of lack of real historical data for this example, we create a "pilot" clinical trial data set by randomly draw a subset from the original data set and use it to construct a set of informative priors for the three Emax model parameters.

A random sample of 5 from the 20 observations for each dose is listed in Table 11.4.

With this "pilot" data set, we fit the Emax model and the estimated parameters are $E_0 = 0.178$ [standard error (SE) = 0.267], Emax = 0.975 (SE = 0.347) and ED50 = 0.063 (SE = 0.089), respectively. Since this would be a pilot study, so we use the point estimates as the prior means, but we double the SEs as the prior standard deviations, to account for potential future data variability. Now for the prior of ED50, since ED50 is expected to be within the dose range 0 to 1, we would still use a Beta-distribution. To derive a suitable prior for ED50, we make use of the fact the if X follows a Beta-distribution with shape parameters α and β, then the mean and variance are $E(X) = \frac{\alpha}{\alpha+\beta}$ and $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Given the estimated mean ED50 = 0.063 and 2 × SE = 0.178 as the standard deviation, we can solve the mean and variance equations to obtain an informative prior for ED50 with $\alpha = 0.054$ and $\beta = 0.810$. This leads to an informative prior set as in Table 11.5.

With this set of informative priors, we can refit the Bayesian model using the following R code chunk:

```
## Define Normal priors for E0 and Emax
prior2 <- list(norm=c(0.178,0.533),norm=c(0.975,0.695),
           beta=c(0,1,0.055,0.810))
## now fit the Emax model again with Bayesian MCMC
b2Mod <- bFitMod(dose, estMean, S, model = "emax",
       nSim = 10000, prior = prior2)
## Print the summary information
b2Mod
```

Table 11.4 Random sample of 5 observations from each dose

Dose	Response				
0.00	0.137	0.358	0.623	−0.308	0.336
0.05	1.354	−0.071	0.584	−0.363	0.339
0.20	1.853	0.996	1.233	1.868	1.011
0.60	0.310	0.310	0.279	0.570	1.604
1	2.174	1.255	0.887	0.425	1.129

Table 11.5 Informative priors for the Emax model

Parameter	Informative prior distribution
E_0	Normal with mean 0.178 and standard deviation 0.533
E_{max}	Normal with mean 0.975 and standard deviation 0.685
ED_{50}	Beta-distribution with $\alpha = 0.054$ and $\beta = 0.810$

This would give the summary information for the 10,000 MCMC as follow:

Summary of posterior draws							
	mean	sdev	2.5%	25%	50%	75%	97.5%
e0	0.331	0.135	5.01e-02	0.2431	0.335	0.423	0.584
eMax	0.830	0.284	3.22e-01	0.6340	0.810	1.004	1.441
ed50	0.294	0.274	3.57e-07	0.0764	0.196	0.450	0.948

As an alternative for the informative priors, one can use *t*-distribution with three degrees of freedoms for the E0 and Emax parameter, as recommended by Novick et al. (2017). It can similarly be implemented as follows:

```
## Define prior for Eo and Emax with t-distribution
Prior3 <- list(t=c(0.178,0.533,3),t=c(0.975,0.695,3),
          beta=c(0,1,0.055,0.810))
## now fit an emax model
b3Mod <- bFitMod(dose, estMean, S, model = "emax",
          nSim = 10000, prior = prior3)
## summary information
b3Mod
```

Summary of posterior draws							
	mean	sdev	2.5%	25%	50%	75%	97.5%
e0	0.332	0.136	5.46e-02	0.2435	0.337	0.425	0.582
eMax	0.783	0.286	2.78e-01	0.5834	0.767	0.959	1.396
ed50	0.257	0.271	5.04e-21	0.0406	0.158	0.396	0.934

As can be seen, essentially the two sets of informative priors produced similar results.

We now summarize these Emax modeling results in Table 11.6.

We emphasize that this example is solely for illustrating how to implement Bayesian Emax modeling using R, rather than making comparisons between the choice of priors or versus the frequentist approach. Furthermore, we did not include

Table 11.6 `Emax modeling example result summary`

Method	Priors	E0 mean (SE), equal tail 95% interval	Emax mean (SE), equal tail 95% interval	ED50 mean (SE), equal tail 95% interval
Frequentist	N/A	0.322 (0.152), (0.024, 0.620)	0.746 (0.236), (0.283, 1.209)	0.142 (0.180), (0, 0.500)[a]
Bayesian	Non-informative	0.386 (0.132) (0.110, 0.629)	0.931 (0.338) (0.342, 1.680)	0.513 (0.326), (0.033, 0.997)
Bayesian	Normal informative base on pilot data	0.331 (0.135) (0.050, 0.584)	0.830 (0.284) (0.322, 1.441)	0.294 (0.274), (0.000, 0.948)
Bayesian	T informative based on pilot data	0.332 (0.136) (0.055, 0.582)	0.783 (0.286) (0.278, 1.396)	0.257 (0.271), (0.000, 0.934)

[a]Adjusted due to ED50 needs to be greater than 0

any Bayesian modeling diagnostics for this example, which would be essential for any kind of Bayesian modeling analyses. We refer the readers to Carlin and Louis (2008) and Gelman et al. (2013) for Bayesian diagnostics topics.

Nonetheless, we can still observe these results by using different methods for the same example data. For the E0 estimates, they are very similar, perhaps stemming from the fact it is a linear term in the model. For the Emax estimates, the Bayesian estimates tend to have bigger variability and this might be due to the use of priors that also bring in bigger variability. For the ED50 estimates, they are the most variable among the 3 parameters, which can be confusing as to which estimate is more plausible. It is noted that the original data set are very noisy in that each dose has a wide range of responses. For example, the placebo responses range from −0.308 to 1.561, whereas the highest dose group responses range from −1.113 to 2.248. As a result, any random sample for a subset of data (which we used to construct informative priors) can be a biased sample for the whole data set and also any prior used for the Bayesian approach can be sensitive to the results.

We also note that in selecting a dose for future development, there are various considerations including efficacy, safety, drug formulation, existing medications in the market, and so forth. Therefore it is important to recognize that the dose-response modeling results provide useful information for dose selection, in the context of the totality of information about the compound being developed.

11.6.3 Summary

In this chapter we covered the basic concepts of the Bayesian approach by using a simple example to illustrate the Bayesian concept, update, and inference. We also applied the approach to phase-II clinical trials by continuing the dose-finding example from Chap. 9. We then summarized and discussed the example results.

Almost all clinical developments have existing historical data that can be used for constructing Bayesian priors. In cases where historical data are rich and relevant, more informative or less diffused priors can be derived and clinical trials can potentially benefit from using them. In cases where historical data are not as rich, less informative or more diffused priors can be used and the Bayesian approach still applies.

References

Berger, J.O. (1985). *Statistical decision theory and bayesian analysis* (2rd ed.). Springer.

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

Brooks, S., Gelman, A., Jones, L., & Meng, X. (2011). Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis* (3rd ed.). Chapman & Hall/CRC.

Chen, D. G., & Peace, K. E. (2011). *Clinical trial data analysis using R*. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Gamerman, D., & Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference* (2nd ed.). Chapman & Hall/CRC Texts in Statistical Science.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC Texts in Statistical Science.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of American Statistical Association, 91*(435), 1343–1370.

Metropolis, N. (1987). The beginning of the Monte Carlo method. Los Alamos Science (1987 Special Issue dedicated to Stanislaw Ulam), 125–130.

Novick, S., Ho, S., & Best, N. (2017). Data-driven prior distributions for a phase-2 COPD dose-finding clinical trial. (submitted for publication).

Oakley J. E., & O'Hagan, A. (2010). SHELF: The sheffield elicitation framework (version 3.0). UK: School of Mathematics and Statistics, University of Sheffield. http://tonyohagan.co.uk/shelf

Zhu, M., & Lu, A. Y. (2004). The counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistical Education, 12*(2), 1–10.

Chapter 12
Overview of Phase III Clinical Trials

12.1 Introduction

After the success of Phase II development, it comes to the Phase III development. Phase III has fundamental objectives in confirming clinical efficacy and investigating no unacceptable safety concerns. Clinical efficacy generally includes effects that are measurable as reflected by clinical outcome assessments such as patient-reported and clinician-reported outcomes and objective measures like survival or lab data. Unacceptable safety concerns generally refer to safety issues that are worse than or new to those of placebo (for lack of standard treatments) or the standard of care (with established treatments).

As this book is focused on Phase II development, details of Phase III development will not be covered. Instead, in this chapter, we highlight the scope of Phase III plans in order to put Phase II trials in a larger context. The discussion that follows includes drug label and target product profile, non-inferiority trial designs, dose selection, drug formulations and patient populations, number of required trials for a labeling claim, number of primary efficacy endpoints, missing data issues, clinical outcome assessments, multi-regional trial issues, and the trend towards development of personalized or precision medicines.

12.2 Scope of Phase III Plans

Phase III developments generally include primary indications for the major patient populations such as adults and minor patient populations including but not limited to pediatrics, targeted at New Drug Application (NDA) in the US, Market Authorization Application (MAA) in Europe, and other regional or country-specific

N. Ting et al., *Phase II Clinical Development of New Drugs*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-4194-5_12

submissions. These developments may continue to various extensions such as different formulations on different devices and other related indications and labeling changes. Generally Phase III also has two sub-phases. Before the first (and usually primary) approved indication, it is called Phase IIIA; after the first approved indication, it is called Phase IIIB. Therefore, a Phase III program can span from a few to many years and cover a large number of patients.

The current requirement is that regulatory agencies ask for the overall Phase IIIA plan to at least include relevant pediatric populations if applicable. Furthermore, with the International Conference on Harmonization (ICH) agreement of using the Common Technical Document (CTD) and its electronic version eCTD, most major international submissions will need to be considered for the Phase IIIA development in order to avoid issues at the time of submissions. As a result, Phase IIIA plan needs to be well thought through and to receive major regulatory agencies' agreement in advance, before its implementation.

12.3 Drug Label and Target Product Profile

Phase III trials are designed to support label claims for the investigational drug. A drug label is an instruction to guide physicians how to prescribe the drug and inform patients regarding appropriate drug use and potential adverse effects. It contains all of the clinical, non-clinical, and manufacturing information and provides detailed drug characteristics and describes how the drug should be dosed. Before approval it must be agreed and finalized between the regulatory agency and sponsor who develops the drug. A complete label usually includes the following sections: Indications and Usage, Dosage and Administration, Dosage Forms and Strengths, Contraindications, Warnings and Precautions, Adverse Reactions, Drug Interactions, Use in Specific Populations, Overdosage, Description, Clinical Pharmacology, Nonclinical Toxicology, Clinical Studies, How Supplied/Storage and Handling, and Patient Counseling Information.

A drug label is usually sketched early in the development stages and used to guide drug development, because it specifies the anticipated or desired characteristics of the new drug. In planning a clinical development program, this draft label is also used to build a Target Product Profile (TPP), which provides the guiding principles for Phase III clinical trial designs. When designing Phase III clinical trials, it is crucial to consider what data to collect in order to support this TPP—in other words, how to enable the drug to be approved with desirable label claims and also reimbursable.

12.4 Phase III Non-inferiority Trial Designs

It is not uncommon that a Phase III clinical trial be designed to compare against an active control, a drug that has been approved and available for patients. This active control may well be the market leader for the same indication for purposes like reimbursement or pricing. As such, the primary hypothesis can be non-inferiority, instead of superiority. The concept of non-inferiority is that from the primary efficacy comparison, the test drug is as good as the active comparator. However, the test drug could be better than the active control in other aspects such as having a better safety profile, easier to use, or more cost-effective. Superiority trial designs are more common and relatively straightforward. In this section, we focus on the non-inferiority trial designs and highlight some of the key issues.

As noted in ICH E10, the primary objective of an active-controlled trial can be either to establish superiority or non-inferiority. The superiority objective is to demonstrate that the study drug is statistically better than the control. Suppose that for this underlying condition, a larger value of the primary measurement implies improvement. Then the one-sided hypothesis testing can be setup as

$$H_0 : \mu_A - \mu_T = 0 \quad \text{versus} \quad H_1 : \mu_A - \mu_T < 0,$$

where μ_A is the mean of the active control and μ_T is the mean of the test drug group.

On the other hand, a non-inferiority objective (ICH E10 2000) is to show that the test drug is as good as or not inferior to the control. The active control has already been demonstrated its efficacy and safety and approved by regulatory agencies. The study drug is being developed to provide comparable efficacy and tolerance with other potential advantages such as having a better safety profile, easier to use, or more cost-effective.

The US Food and Drug Administration (FDA) *Guidance for Industry* on Non-inferiority Trials (2010) provides guidance to ensure that a non-inferiority conclusion not only pertains to study treatment being non-inferior to the active control, but also being superior to placebo. We adopt the guidance and highlight some key considerations for non-inferiority trial designs.

Let M_1 denote the entire effect the active control has relative to placebo, and M_2 denote the largest loss of effect that would still be clinically acceptable for the test drug. Essentially M_2 is the non-inferiority margin to be used for establishing that the test drug is no worse than the active control by M_2 difference in effect.

M_1 and M_2 can be more formally defined as:

M_1 –the entire effect of the active control assumed to be present in the non-inferiority trial

M_2 –the largest clinically acceptable difference (degree of inferiority) of the test drug compared to the active control

Then the non-inferiority hypothesis testing is set up as:

$$H_0 : \mu_A - \mu_T \geq M_2 \quad \text{versus} \quad H_1 : \mu_A - \mu_T < M_2.$$

The M_1 effect is determined from historical information (not from the current trial), and the M_2 can be calculated as a difference from or a percentage of M_1. In principle, M_2 needs to be clinically acceptable and agreeable by the regulatory agency. Only after the agreement is reached, such a margin is used for the Phase III trial design and upon which sample size calculations can be based.

Furthermore, the non-inferiority trial must demonstrate assay sensitivity which is the ability to distinguish an effective treatment from an ineffective one, in order to conclude non-inferiority. Assay sensitivity can be demonstrated directly from the trial when placebo is present, otherwise indirectly from historical information as described in the guidance "…by reference to results in previous placebo-controlled studies of the active control drug."

12.5 Dose and Regimen Selection, Drug Formulation and Patient Populations

12.5.1 Dose and Regimen Selection

The choice of doses and regimens for a Phase III design follows the guidance of the draft label or TPP. If the TPP dictates a single dose for approval and marketing, then the strategy of this Phase III clinical program should be centered on the objective of delivering a single efficacious and safe dose for long-term use. In certain development programs, the project team devotes good efforts in Phase II to obtain clear dosing information. At the End-of-Phase II (EoPII) meeting with FDA or Scientific Advice (SA) meeting with European Medicinal Agency (EMA), the sponsor team should reach an explicit agreement with the Agency regarding dose(s) to be used for Phase III development of the candidate intervention. The objective is that, afterwards, the choice of dose(s) could be made simple and clear to satisfy regulatory requirements.

However, in most of the development programs for interventions treating chronic diseases, the dosing information from Phase II is not as clear. Usually there are ample discussions during the EoPII or SA meeting between the sponsor and FDA or EMA. In cases where it is ethical to use a placebo control for the indication, then one recommendation is to consider three test doses against placebo, using a gate-keeping multiple comparison adjustment (if monotonicity can be safely assumed) to avoid inflation of false positives. Rationale of this strategy is clarified in the next few paragraphs.

The main trial purpose is to demonstrate the target dose is superior to placebo, and maintains an acceptable long term safety profile. It helps to allocate a dose

higher than the target dose so that in case the target dose does not provide sufficient efficacy, this higher dose can serve as a backup, because efficacy in Phase III trials tends to be diluted from efficacy in Phase II trials. There are several reasons for the tendency of lower efficacy in Phase III trials. One of the likely reasons is the patient selection criteria—Phase II tends to allow the test drug an optimal opportunity to demonstrate its efficacy. As such, the inclusion/exclusion criteria in Phase II tend to be more restrictive. If the patient's disease is too severe, it may be difficult for any drug to improve disease symptoms. If the patient's disease is too mild, then the efficacy from test drug can be difficult to differentiate from placebo.

Hence, in general, Phase II results may reflect the "best cases" of the efficacy observed from the test drug. The inclusion/exclusion criteria for Phase III trials are more relaxed, with a hope to recruit a wider patient population which is similar to the real world patient population. Another reason is that Phase III trials are longer-term trials, while Phase II tends to be shorter term. Efficacy observed from shorter-term trials in Phase II may not be able to translate into longer-term efficacy in Phase III. One situation is that for patients with chronic diseases, their expectation changes over time—after initial symptom improvement, they tend to take this improved state for granted, and anticipate further improvement of their conditions. Given this response shift from patients, it helps to include a higher dose in Phase III so that if the target dose does not deliver the anticipated long term efficacy, a higher dose may be able to.

Similarly, a lower dose may be helpful in case the target dose tends to be less safe after longer term follow up. One of the unexpected risks in Phase III development is the late developed adverse events (AE). Because Phase I and II drug exposure are shorter term, most of the AE's could have been observed at the early stage of drug development. In general, even for a relatively safe new drug, patients may react within the first few weeks of drug exposure, because the human body is not used to a new chemical or biological agent in their system. For most drugs, these AE's developed at early stage may disappear after repeated exposure, because the human body adapts this new agent over time. However, some toxicity may be related to drug accumulation in the body. Adverse events caused by this type of drug exposure may not be detected from short-term trials. Then in Phase III, with longer term accumulation, this kind of AE's are observed, and it will be very difficult to manage. As a way to avoid or mitigate such safety risk, it helps to allocate a lower dose in Phase III so that if the target dose is not safe, then this lower dose could serve as a backup.

For most of the new drugs, monotonic efficacy dose-response relationship can be expected. In fact, this relationship should have been observed from Phase II. On this basis, gate-keeping procedure for multiple comparison adjustment is very sensible. However, in case it is not known whether the efficacy dose-response is monotonic, then the preferred method for multiple comparison method would be non-gate-keeping procedures such as the Hochberg procedure (Hochberg 1988).

In case for the underlying indication, if it is not ethical to use placebo control, then an active control should be used for Phase III development. At this stage, it is critical for the project team to discuss whether superiority trials or non-inferiority

trials will be designed for Phase III. If the choice is a superiority trial, then the previous discussions regarding placebo-controlled considerations can still be applicable. However, if the decision is to design non-inferiority trials, then the team should be clear about the advantages of this study drug. If the efficacy is only non-inferior to the active control, then what will be the benefit to develop this new drug candidate? Only after the answer to this question can be justified, then the Phase III development program can be reasonable.

For non-inferiority designs, one of the first questions would be which active control should be used. In some indications, the approved market comparators can be different depending on the countries where the clinical trial will be conducted. It is also necessary to consider the submission strategy of this new drug, which regulatory agencies will the new drug be submitted to and in what sequence? The EMA and FDA considerations could be different for this indication. Also, various Asian or South American countries may have different requirements. Hence it is critical for the project team to determine a submission strategy first, for example, to decide which country or region the new drug will be submitted first and where the Phase III trials will be conducted. Then, active control should be selected according to these considerations.

After the active control is determined, it is also important to consider the dosage/regimen of the active control agent. Which dose(s) or regimen(s) of this active agent will be used in the Phase III comparisons? After some of these major decisions, then the challenge will be selection of the primary endpoint and the non-inferiority margin to be applied to this endpoint, which will be a joint decision between the sponsor (e.g., pharmaceutical or biotech company) and the regulatory agencies (often FDA or EMA). Only after these issues are settled can sample sizes be calculated meaningfully. Again, even under the non-inferiority setting, it is useful to consider multiple doses of test drug in comparison with the active control.

If the TPP indicates that multiple doses (or regimens) will be marketed, then the potential range of doses or regimens will need to be studied. Again, for Phase III trials, it will be helpful to design a higher dose above the targeted upper dose range and a lower dose below the planned lower limit of dose range. This strategy will help reduce the risks that the long-term efficacy of the high dose is not sufficient or that the safety profile of the low dose is worse than anticipated.

12.5.2 Drug Formulations

In the early phase and Phase II developments, the drug formulations (including delivery devices) may not be the one used in the final product. Nevertheless, in Phase III trials, the final formulation should be used in order to assure trial results reflect the product to be used by patients, after the product is approved by regulatory agencies. In certain situations, if final formulation is not used in Phase III

trials, some bridging work will likely be required by regulatory agencies. Therefore, it is necessary to have proactive communications with regulatory agencies, before designing such Phase III trials.

12.5.3 Patient Populations

Similar to drug formulations, in the early phase and phase II developments, the populations studied are not necessarily the final targeted treatment populations. Nonetheless, in Phase III clinical trials, it is important to investigate the patient population to be treated by the medication, with the expectation that treatment effects observed in Phase III trials can reflect the responses from the indicated patient population. Furthermore, if successful, the trial results will be included in the drug label to guide physicians and patients. In certain situations, if the targeted population (e.g., infants) is not used in Phase III trials, medical justification will likely be required by regulatory agencies. Therefore it is necessary to have proactive communications with regulatory agencies before designing such Phase III trials.

12.6 Number of Phase III Trials for a Labeling Claim

Traditionally, regulatory agencies require two pivotal Phase III clinical trials to demonstrate replicate evidence for a labeling claim. For example, in order to demonstrate a corticosteroids is efficacious in improving lung function in mild-to-moderate asthma patients, two replicated clinical trials need to demonstrate that the drug improves patients' lung function by a clinically meaningful and statistically significant magnitude (typically at two-sided 5% alpha level for each study).

It is stated in ICH E9 that the evidence from a single trial can be considered sufficient in some circumstances (ICH 1998b). Examples may be cases where benefits on mortality are seen or where the disease is difficult to study. Additional potential acceptance is expressed in the CPMP Points to Consider document labeled "Points to Consider on Application with 1. Meta Analyses; 2. One Pivotal Study" (EMEA 2001). Therefore, in recent years, sponsors have proposed to conduct single pivotal Phase III clinical trial for confirmatory and submission purposes.

Technically for a single pivotal trial to replace two replicated trials, the statistical significance level will need to be at two-sided 0.125%, much smaller than the usual two-sided 5% level for two replicated trials. The reason is that in terms of the risk of making a type I error in evaluating a null hypothesis of equality between a test and a control treatment, only rejection in favor of the test over the control is used for approval ($0.00125 = 2 \times 0.025^2$). Using one-sided testing, requiring two trials with

p-values <0.025 is equivalent to requiring one trial with a *p*-value <0.000625, which is the same standard as requiring a *p*-value <0.125% for two-sided testing.

In terms of power, a single trial with the same sample size as the total sample size from two replicated trials will have greater power. For example in the case where two independent trials are each powered at 90% for two-sided alpha = 0.05, yielding an overall power of 81% ($=0.90^2$), a single trial using two-sided alpha = 0.00125 would preserve the 90% power. However, it is noted that replication of evidence can be more important than a smaller *p*-value. Two trials expose the study drug to two independent cohorts of patients and provide two separate pieces of evidence. The readers are referred to Shun et al. (2005) for detail discussions of these and other related issues.

There has been some precedent that FDA and EMA accept single pivotal Phase III trial for a two-sided significance level of 1%, for some Phase III trials that fit the ICH and CPMP guidance, as a balance among unmet medical needs, controlling the overall type I error rate and the potential power increase from the single pivotal trial. It is noted that such an agreement will need to be obtained between the sponsor and the regulatory agency, before the trial is designed and conducted.

12.7 Number of Primary Efficacy Endpoints

Depending on the product indication and treatment effects, a single primary efficacy endpoint or multiple primary efficacy endpoints can be used for Phase III pivotal trials. For example, a combination of inhaled corticosteroids and long-acting beta agonist for asthma, typically require more than one primary efficacy endpoint as this combination drug needs to demonstrate treatment effects from each of its components. For this example, a longer term treatment effect endpoint such as pre-dose FEV_1 (forced expiratory volume in one second, essentially the maximal amount of air one can forcefully exhale in one second) after six weeks of treatment can be used to demonstrate the corticosteroids anti-inflammatory effect. On the other hand, a quick response such as two-hour post dose FEV_1 on the first treatment day can be used to demonstrate the beta agonist bronchodilator effect.

Co-primary or multiple primary efficacy endpoints can be useful and even necessary for some Phase III trials, primarily due to two reasons. One reason is that the efficacy can be better supported by more than one primary measure. For example, the asthma patients can benefit from both lung function improvement measured by FEV_1 and the asthma symptoms measured by a composite score of coughing, wheezing, shortness of breath, chest tightness, and so forth. The second reason is that different regulatory agencies, reflected by their regional medical practice, may require a different primary efficacy endpoint. For example, the FDA tends to focus more on the lung function whereas the EMA tends to focus on symptoms.

Note that when there is more than one primary efficacy analysis, such as using co-primary endpoints, or more than one primary time point, there are multiplicity issues. In Phase III trials, it is necessary to consider multiplicity adjustments or management methods to control the overall type I error rate for the primary analyses, as the results will be used for labeling claims. Generally speaking, multiplicity adjustment strategies include sequential gate-keeping comparisons without the need to adjust *p*-values, parallel or simultaneous comparisons using adjusted *p*-values, or a combination of the two strategies. There are various multiplicity adjustment methods that are suitable for different situation and needs. Please refer to Westfall et al. (2011), Dmitrienko and D'Agostino (2013), and Dmitrienko et al. (2013) for more details on this topic.

12.8 Missing Data Issues

Missing data can influence clinical trial results, which may have profound impact on drug developments. The best way to address missing data is at the design and conduct stages, rather than at the analysis stage.

Elements in the design stage and in the conduct stage of clinical trials can help to limit missing data (National Research Council 2010; Little et al. 2012a, b). At the design stage, the following eight ideas have been suggested: (1) target a population that is not adequately served by treatment, (2) include a run-in period, (3) allow a flexible treatment regimen, (4) consider an add-on design, (5) shorten the follow-up period, (6) allow the use of rescue medications, (7) consider a randomized withdrawal design (for long-term efficacy), and (8) avoid outcome measures that are likely to lead to substantive missing data. At the conduct stage, the following eight ideas have been suggested: (1) select investigators with a good record for collecting complete data, (2) set acceptable target rates for missing rates and monitor the progress made, (3) provide monetary and nonmonetary incentives to investigators and participants, (4) limit the burden and inconvenience of data collection, (5) provide continued access to effective treatments after the trial (before treatment approval), (6) train investigators and study staff on ways to keep participants in the study until the end, (7) collect information from participants regarding likelihood that they will drop out, and (8) keep contact information for participants up-to-date. The relevance of the suggestions at the design and conduct stages varies greatly according to setting, and they may have limitations or drawbacks that need to be considered.

At the analytic stage, there are at least four approaches to address the missing data problem, which are applicable to patient-reported outcomes as well as other types of outcomes (Fairclough 2010). One approach is to remove patients with missing or incomplete forms from the analysis and only analyze complete cases. A second approach is to impute the missing data. A third approach to address the problem of missing data is through the application of a likelihood-based approach using repeated-measures models or mixed-effect models (Mallinckrodt et al. 2008);

in this approach, every subject would contribute his or her available (observed) measurements. The fourth approach is especially relevant when missing data are not missing at random (MAR) and hence depend on the (unknown) missing value, when missing data are said to be non-ignorable. The advantages and disadvantages of these approaches—along with a guidance on how to handle missing data—are found in National Research Council (2010), and Little and Rubin (2002).

For example, at the analysis stage, alternative approaches may be used depending on the *missingness* mechanism. When data are MAR, multiple imputation (Carpenter and Kenward 2013) and mixed-effect model with repeated measures (Fairclough 2010; Fitzmaurice et al. 2011) are routinely used. In certain situations, the missingness mechanism may be obvious by inspecting the study design or other aspects of the study conduct. However, in general, it may not be possible to ascertain whether the assumption of MAR can be justified. As a best practice, it is therefore advisable to perform sensitivity analysis to ensure the robustness of the findings under alternate scenarios (Little and Rubin 2002; National Research Council 2010). Pattern mixture models are sometimes used in sensitivity analysis, with missing values imputed under a plausible scenario in which the missing data are considered missing not at random (MNAR) (Little and Rubin 2002). An alternative approach, which is relatively less dependent on assumptions, involves searching for a *tipping point* that reverses the study conclusion (O'Kelly and Ratitch 2014).

The last observation carried forward (LOCF) analysis has been used in many drug approvals over the years (Ting 2000). More recently there has been increased discussion about the appropriate use of LOCF (Akacha et al. 2015; Mallinckrodt et al. 2013a, b).

12.9 Phase III Clinical Outcome Assessments

There are various types of primary efficacy endpoints for Phase III trials, depending on trial objectives, disease nature, measurement sensitivity, medical practice, regulatory requirements and so forth. The key consideration is for the primary efficacy endpoints to accurately reflect clinical outcomes from study treatments, so that the outcome assessments can help confirm clinical efficacy to obtain regulatory approval.

Chapter 1 highlights clinical outcome assessments (patient-reported outcomes, clinician-reported outcomes, observer-reported outcomes, performance-based outcomes), which also included patient-reported health-related quality of life. They are often used in Phase III trials, as well as in Phase II trials, as primary endpoints or secondary endpoints or both. The formal validation of a clinical outcome assessment (COA) instrument should begin in early development, much before Phase III begins, with content validity. Content validity is the extent to which an instrument covers the important concepts of the unobservable or latent attribute (e.g., depression, anxiety, physical functioning, self-esteem) the instrument purports to

measure. It is the degree to which the content of a measurement instrument is an adequate reflection of the construct (the same "thing" or concept) to be measured. Steps to ensure content validity are found in Patrick et al. (2011a, b).

After content validity has been presumably established, the selection of a COA instrument must also consider its measurement or psychometric properties, again before Phase III trials begin such as in Phase II trials and pre-Phase III non-interventional, methodological studies. Phase III trials can be used later to confirm the measurement properties of a COA. While content validity is based on qualitative methods, measurement properties of an instrument are grounded in quantitative analysis.

The selected instrument must be psychometrically sound. Is the instrument measuring what it is intended to measure—is it valid? Does it give accurate measurements—is it reliable? Measurement characteristics including reliability and validity are fundamental aspects for judging the quality and merits of a COA instrument. Details on assessing reliability and validity can be found elsewhere and involve correlations, means and regression methods, as well as theoretical expectations (Fayers and Machin 2016; Streiner et al. 2015; Cappelleri et al. 2013; de Vet et al. 2011).

Assuming that a COA has sufficient content validity and psychometric properties, it can be considered as a primary endpoint or a secondary endpoint for a label claim or simply to expand the base of knowledge on patient-centered outcomes. The number and timing of COAs are influenced by the study objectives, such as when meaningful change is expected, and practical considerations, such as patient burden. Details on key considerations in the design of a longitudinal study for COAs can be found in Fairclough (2004, 2005, 2010).

Consider one type of COA: a patient-reported outcome (PRO). When the objective of a study is to compare a PRO in subjects who experience the same type of condition during a given phase of treatment, assessments can be planned at times when clinically relevant events are expected to occur or at times that correspond to a distinct, meaningful phase of the intervention or disease. Such assessment is more common for a design with a relatively short duration. Many variations exist. Among them, for example, is when differences in PRO values are expected during only the early period of therapy. A breast cancer trial of adjuvant therapy in which a 16-week dose-intensive regimen was compared with a more traditional 24-week regimen is an illustration of such an event-driven design (Fetting et al. 1998). Three assessments were planned—prior to (baseline assessment), during, and after therapy—where each phase of the disease or its treatment was considered distinct with respect to the PRO of interest.

In event-driven designs where each assessment is conceptually identified with a landmark event, repeated-measures models for longitudinal data (with time taken as a categorical covariate) are an appropriate choice. Note that assessments for all subjects should be taken at the same points in time (e.g., week 6, week 10, and week 24), where points in time need not be equally spaced. Repeated-measures models may also be useful in some studies with only a few assessments.

When the scientific questions involve a more extended period, or when the phases of the disease or its treatment are not distinct, the longitudinal designs are based on or driven by time (Fairclough 2005, 2010). These designs are appropriate for chronic conditions where therapies are given over elongated periods, such as diabetes and arthritis.

In time-driven designs the duration of therapy may be indeterminate at study onset, with therapy intended to be given to a patient until it is not efficacious or produces unacceptable toxicity. For instance, patients with advanced renal cell carcinoma were randomized to receive either repeated 6-week cycles of sunitinib (experimental) or interferon alfa (control) (Cella et al. 2008). Doses were adjusted in response to symptoms of toxicity. Treatment in both groups was continued until the occurrence of death, unacceptable adverse events, or withdrawal of consent. Patients were asked to complete the PRO questionnaires before any clinical activities during visits to the study clinics at screening, on days 1 and 28 of each 42-day treatment cycle, and at the end of treatment or study withdrawal.

Time-driven designs are associated with mixed-effect models for studies where time is often conceptualized and taken as a continuous variable. Mixed-effect models are useful when the timing of assessment differs widely among individuals, studies have a large number of PRO assessments, or changes over time are to be modeled with a smaller number of parameters than that required for a repeated-measures model (with time as a categorical covariate).

A major factor when deciding on the timing of the COA assessment, both initially and subsequently, is the recall period of the COA questionnaire. Because individuals have better recall for major events and more recent experiences, the period of accurate recall for measuring certain areas (e.g., erectile dysfunction, physical well-being) is between one and four weeks, whereas the period of recall for the frequency and severity of symptoms (e.g., pain, fatigue) is accurate over shorter periods such as at the time of patient completion of the COA or the past 24 h. That said, it should be noted that recall period established by the developers of the COA instrument should be used.

For regulatory claims on a COA, the recall period with the shortest time frame consistent with the instrument's purpose or intended use (e.g., when feasible, a recall period referenced to the patient's current or recent state) is preferable to a recall period that is based on a longer period, a comparison of a patient's current state with an earlier period, and a self-reported average over time (FDA 2009). The frequency of the assessments on COAs should be frequent enough to capture meaningful change over a sufficient duration but not frequent enough to cause excessive burden on participants.

Patients who drop out of a study prematurely are generally more likely to have a less favorable score on a COA because of side effects or no effect of treatment. A treatment arm with a high rate of dropout is likely to give an artificially more favorable outcome because only the healthiest of the patients remain on treatment, leading to selection bias and overly optimistic estimates of treatment effect. It is therefore desirable to have a COA assessment in conjunction with premature withdrawal from the study. If the research objective extends to off-therapy

assessments, then they can be made by continuing the COA assessments after discontinuation. The off-therapy assessments can always be excluded if deemed uninformative or irrelevant to the research question. Including the off-therapy assessments after discontinuation allows them to be available, should they be determined to be of interest.

Missing data on COAs can occur as missing items or missing questionnaires. Missing items involve the lack of responses for some specific items; missing questionnaire involves patients who may fail to complete and return the whole questionnaire. Many instruments include well-documented procedures by their developers on how to handle missing items. Such recommendations by developers are typically the preferred way to address missing items. Missing questionnaires are a more complex situation than missing items. Missing questionnaires can happen as a result of dropout from the study or randomly failing to fill out an entire questionnaire. Section 12.8, which covers this topic generally, is also applicable to missing data on COA questionnaires.

12.10 Multi-regional Phase III Clinical Trial Issues

Phase III clinical trials tend to be larger in sample size and targeted for multiple international submissions for approval. Globalizing clinical trials may enable more efficient drug and product development, which would then enhance product affordability and foster further investment in research and development. As a result, they tend to be multi-regional and may cover the USA, Europe, China, Japan, other Asian-Pacific countries, South America, or African countries. Different countries and regions have different medical practice and regulatory requirements. Therefore, multi-regional clinical trials (MRCT) will have to be harmonized, which prompts actions from the FDA, EMA and Japanese Pharmaceuticals and Medical Devices Agency (PMDA) to release guidelines. International harmonization will promote to conduct the MRCT more appropriately and more efficiently on drug development, as well as to avoid duplicative developments.

For example, Japan in their Basic Principles on Global Clinical Trials (2007) document requires that "A global trial should be designed so that consistency can be obtained between results from the entire population and the Japanese population." It also recommends two methods of calculating Japanese sample size, which generated debates and discussions, as seen, for example in Ikeda and Bretz (2010).

Despite the globalizations of MRCTs, there are emerging challenges in the conduct of such trials at different regions regarding potential regional heterogeneity. This heterogeneity, which needs to be carefully considered, is generated by different requirements from different regulatory agencies in different countries with regard to the study design and conduct of MRCTs. As reviewed by Girman et al. (2011), there are (1) different endpoints required by different regions since different regulatory authorities have different standards for primary, co-primary, or key secondary endpoints; (2) different time points required for hypothesis testing since the

expected treatment duration or the primary time point for measuring treatment response in a trial may not be consistent across regions; (3) different experimental designs which are modified to match the different characteristics from different regions in a MRCT; (4) different non-inferiority margins required by different regulatory agencies; and (5) different analytic patient populations defined in analysis plan by different regions.

Any of these aforementioned differences could lead to inflated type I and type II errors from the perspectives of statistical analysis and final recommendations. Therefore, attempts and attentions should be made to resolve these differences prior to MRCT's initiation. It then becomes imperative to enforce and follow the guidance from global harmonization from different health and regulatory authorities as in International Conference on Harmonization (http://www.ich.org). In addition to regulatory considerations, regional differences can also be thought of in terms of intrinsic or extrinsic factors (ICH E5 1998a). Intrinsic factors may be race or genetic, and extrinsic factors could include diet, smoking or cultural related factors. Currently ICH is drafting E17, which addresses MRCT issues.

12.11 The Trend Towards Personalized or Precision Medicines

On April 14, 2003 the US National Human Genome Research Institute, the Department of Energy, and their partners in the International Human Genome Sequencing Consortium announced the essential completion of a high-quality version of the human sequence, including the creation of physical and genetic maps of the human genome. This major milestone enhanced the possibility of personalized or precision medicine, which involves getting the right therapy to the right patient at the right time. Depending on each individual's genetic content or other molecular or cellular analysis, it is possible to identify individuals who may benefit the most from certain medicine under a customized dosing regimen. Advances in personalized medicine will create a more unified treatment approach specific to the individual and their genome. Personalized medicine may provide better diagnoses with earlier intervention, and more efficient drug development and therapies.

With this new trend, enrichment strategies will be popular and useful for Phase III trials to identify patient subgroups with certain characteristics such as certain genotyping or phenotyping, to benefit from optimized treatment effects. From the standpoint of improving efficacy, safety and public health, there is a general consensus about the value of targeting treatment at patients that are highly likely to benefit from it. A core belief behind the concept of personalized medicine is that it may expedite drug development and lead to greater efficacies in health care. Examples include selecting patients whose disease does not spontaneously disappear or exhibit a large degree of variability, who are likely to comply with treatment, who are likely to have a high rate of disease progression, or who have

some characteristic that suggests they can respond to the treatment. For more detail, the reader is referred to the draft FDA guidance document on "Enrichment Strategies for Clinical Trials to Support Approval of Human Drug and Biological Products" (2012).

12.12 Summary

In this chapter, we highlight the scope of Phase III plans and various issues associated with the Phase III development. These issues include drug label and target product profile, non-inferiority trial designs, dose selection, drug formulations and patient populations, number of required trials for a labeling claim, number of primary efficacy endpoints, missing data issues, clinical outcome assessments, multi-regional trial issues, and the trend towards development of personalized medicines.

Although there are many other issues on Phase III clinical trials, we believe this chapter provides a good general overview and introduction to the Phase III developments. The reader is encouraged to refer to literature and books that cover in-depth knowledge about Phase III developments.

References

Akacha M., Bretz F., Ohlssen D., Rosenkranz G., & Schmidli H. (2015). *Biopharmaceutical Report* (Vol. 22, No. 3).

Cappelleri, J. C., Zou, K. H., Bushmakin, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. (2013). *Patient-reported outcomes: Measurement, implementation and interpretation*. Boca Raton, Florida: Chapman & Hall/CRC Press.

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. New York: Wiley.

Cella, D., Li, J. Z., Cappelleri, J. C., Bushmakin, A., Charbonneau, C.,Kim, S. T. et al. (2008). Quality of life in patients with metastatic renal cell carcinoma treated with sunitinib versus interferon-alfa: Results from a phase III randomized trial. *Journal of Clinical Oncology, 26*, 3763–3769.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. New York, NY: Cambridge University Press.

Dmitrienko, A., & D'Agostino, R. B. (2013). Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine, 32*(29), 5172.

Dmitrienko, A., D'Agostino, R. B., & Huque, M. (2013). Key multiplicity issues in clinical drug development. *Statistics in Medicine, 32,* 1079–1111.

EMEA European Agency for the Evaluation of Medicinal Products. (2001). *Points to consider on application with 1. Meta-analysis; 2. One pivotal study*.

Fairclough, D. L. (2004). Patient reported outcomes as endpoints in medical research. *Statistical Methods in Medical Research, 13,* 115–138.

Fairclough, DL. (2005). Analysing longitudinal studies of QoL. In P. Fayers & R. Hayes (Eds.), *Assessing quality of life in clinical trials* (p. 149–165). Oxford, England: Oxford University Press.

Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials* (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Fayers, P. M., & Machin, D. (2016). *Quality of life: The assessment analysis and reporting of patient-reported outcomes* (3rd ed.). Chichester, UK: Wiley.

Fetting, J. J., Gray, R., Fairclough, D. L., Smith, T. J., Margolin, K. A., Citron, M. L., et al. (1998). A 16-week multidrug regimen versus cyclophosphamide, doxorubicin and 5-flurouracil as adjuvant therapy for node-positive, receptor negative breast cancer: An intergroup study. *Journal of Clinical Oncology, 16,* 2382–2391.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Hoboken, New Jersey: Wiley.

Food and Drug Administration (FDA). (2012). *Draft guidance for industry enrichment strategies for clinical trials to support approval of human drug and biological products*.

Food and Drug Administration (FDA). (2010). *Draft guidance for industry non-inferiority clinical trials*.

Food and Drug Administration (FDA). (2009). *Guidance for industry patient-reported outcome measures*: *Use in medical product development to support labeling claims*.

Girman, C. J., Ibia, E. I., Menjogo, S., Mak, C., Chen, J., Agarwel, A., et al. (2011). Impact of different regulatory requirements for trial endpoints in multiregional clinical trials. *Drug Information Journal, 45,* 587–594.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika, 75,* 800–802.

International Conference on Harmonization (ICH). (1998a). *Harmonized tripartite guideline: Ethnic factors in the acceptability of foreign clinical data—E5*.

International Conference on Harmonization (ICH). (1998b). *Harmonized tripartite guideline: Choice of control group and related issues in clinical trials—Statistical principles for clinical trials—E9*.

International Conference on Harmonization (ICH). (2000). *Harmonized tripartite guideline: Choice of control group and related issues in clinical trials—E10*.

Ikeda, K., & Bretz, F. (2010). Sample size and proportion of Japanese patients in multi-regional trials. *Pharmaceutical Statistics, 9*(3), 207–216.

Little, R. J., D'Agnostino, R., Cohen, M. L., Didkersin, K., Emerson, S. S., Farrar, J. T., et al. (2012a). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine, 367,* 1355–1360.

Little, R. J., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Neaton, J. D., et al. (2012b). The design and conduct of clinical trials to limit missing data. *Statistics in Medicine, 31,* 3433–3443.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y., & Mancuso, J. P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal, 42,* 303–319.

Mallinckrodt, C., Roger, J., Chuang-Stein, C., Molenberghs, G., Lane, P., O'Kelly, M., et al. (2013a). Missing data: Turning guidance into action. *Statistics in Biopharmaceutical Research, 5*(4), 369–382. doi:10.1080/19466315.2013.848822

Mallinckrodt, C., Roger, J., Chuang-Stein, C., Molenberghs, G., O'Kelly, M., & Ratitch, B., et al. (2013b). Recent Developments in the Prevention and Treatment of Missing Data. *Therapeutic Innovation & Regulatory Science* (2014), *48*, 68, doi:10.1177/2168479013501310

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials*. Washington, DC: National Academies Press (http://www.nap.edu/catalog.php?recordid-12955)

O'Kelly, M., & Ratitch, B. (2014). *Clinical trials with missing data: A guide for practitioners*. Hoboken, NJ: Wiley.

Patrick, D. L., Burke, L. B., Gwaltney, C. H., Kline Leidy, N., Martin, M. L., Molsen, E., et al. (2011a). Content validity—Establishing and reporting the evidence in newly developed patient reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good

research practices task force report: Part 1—Eliciting concepts for a new PRO instrument. *Value in Health, 14*, 967–977.

Patrick, D. L., Burke, L. B., Gwaltney, C. H., Kline Leidy, N., Martin, M. L., Molsen, E., et al. (2011b). Content validity—Establishing and reporting the evidence in newly developed patient reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part 2—Assessing respondent understanding. *Value in Health, 14*, 978–988.

Shun, Z., Chi, E., Durrleman, S., & Fisher, L. (2005). Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger versus two smaller pivotal studies. *Statistics in Medicine, 24,* 1619–1637.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). New York, NY: Oxford University Press.

Ting, N. (2000). Carry-forward analysis. In *Encyclopedia of biopharmaceutical statistics* (pp. 103–109). New York: Marcel Dekker, Inc.

Westfall, P., Tobias, R. D., & Wolfinger, R. D. (2011). *Multiple comparisons and multiple tests using SAS* (2nd ed.). Cary, U.S.A: SAS Institute.