Contemporary Clinical Trials 45 (2015) 13–20

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Overview of multiple testing methodology and recent development in clinical trials

Deli Wang *, Yihan Li, Xin Wang, Xuan Liu, Bo Fu, Yunzhi Lin, Lois Larsen, Walter Offen

Data and Statistical Science, AbbVie Inc., 1 North Waukegan Road, North Chicago, IL 60064-6075, USA

ARTICLE INFO

Article history:
Received 27 April 2015
Received in revised form 17 July 2015
Accepted 19 July 2015
Available online 22 July 2015

Keywords:
Multiplicity
Clinical trial
Gatekeeping procedure
Graphical approach

ABSTRACT

Multiplicity control is an important statistical issue in clinical trials where strong control of the type I error rate is required. Many multiple testing methods have been proposed and applied to address multiplicity issues in clinical trials. This paper provides an application oriented and comprehensive overview of commonly used multiple testing procedures and recent developments in statistical methodology in multiple testing in clinical trials. Commonly used multiple testing procedures are applied to test non-hierarchical hypotheses and gatekeeping procedures can be used to test hierarchically ordered hypotheses while controlling the overall type I error rate. The recently developed graphical approach has the flexibility to integrate hierarchical and non-hierarchical procedures into one framework. A graphical multiple testing procedure with "no-dead-end" provides an opportunity to fully recycle α across hypothesis families. Two hypothetical clinical trial examples are used to illustrate applications of these procedures. The advantages and disadvantages of the different procedures are briefly discussed.

1. Introduction

Multiple testing problems or multiplicity issues are often encountered in modern clinical trials. Multiple objectives are typically designed to be addressed in a single clinical trial. The objectives can be defined by multiple dose levels, multiple endpoints, or multiple populations. The multiplicity issue associated with multiple objectives in clinical trials is one of the most important statistical problems for the pharmaceutical industry and regulatory agencies. From the industry's perspective, early stage clinical trials (such as phase I or II) without proper multiplicity control may result in false positive findings, which could result in advancing ineffective drugs to the confirmatory stage, and wasting resources that could be used to develop alternative effective drugs. From the agencies' perspective, clinical trials where the multiplicity issue is not properly handled may lead to unsubstantiated claims for the effectiveness of a drug as a consequence of an inflated rate of false positive conclusions. EMA published a guidance document "Points to Consider on Multiplicity Issues in Clinical Trials" in 2002 [1]. FDA is expected to issue a draft guidance on multiplicity issues in the near future.

Multiplicity issues are primarily related to controlling the type I error rate, which is the probability of rejecting the null hypothesis when it is true. There are two levels of the type I error rate. One is referred to as the comparison-wise error rate, which is for a single hypothesis. The other one is referred to as the family-wise error rate (FWER), which is for all hypotheses being tested. Regulatory guidance requires strong control of the FWER (or the overall type I error rate) which is the probability of rejecting at least one true null hypothesis under all possible configurations of the null hypotheses. The FWER can be heavily inflated if multiple testing is not properly addressed in clinical trials. As an illustrative example, we assume that there are m independent tests and each test has a type I error rate of 0.05. If no adjustment of the significance level is made for each test, i.e., 0.05 is still used for each hypothesis test, then the final family-wise type I error rate for all m tests will be $1-(1-0.05)^m$, which is inflated compared to the pre-specified 0.05 level. For example, if $m = 2$, the inflated type I error rate will be 0.095; if $m = 3$, the inflated error rate is 0.143. The bigger m is, the higher the inflated type I error rate. Therefore, multiple testing procedures are needed to control the FWER at the desired level. If m is large (e.g., in genetic studies where over 10,000 hypotheses are tested), the control of the false discover rate (FDR) [36] is commonly used. However, FDR may not necessarily control FWER in the strong sense which is typically required in the registrational clinical trial setting.

In clinical trial designs, due to different clinical interpretations and inter-relationship among the hypotheses, we can generally categorize the relationships among multiple hypotheses into two categories: 1) non-hierarchical relationships, and 2) hierarchical relationships. Hierarchical relationships include situations where certain hypotheses (or groups of hypotheses) should only be tested following rejection of other hypotheses. Non-hierarchical relationships are appropriate when strict ordering among hypotheses is absent. Correspondingly, multiple testing procedures are tailored towards either non-hierarchical testing or hierarchical testing. In the following sections, the multiple testing procedures discussed in this paper will be grouped into hierarchical procedures and non-hierarchical procedures.

* Corresponding author.
E-mail address: deli.wang@Abbvie.com (D. Wang).

http://dx.doi.org/10.1016/j.cct.2015.07.014
1551-7144/

Multiple testing methods have been widely investigated in the literature. For example books on multiplicity methodology have been authored by Hochberg and Tamhane [2], Westfall and Young [3], and Hsu [4]. Hommel, Bretz and Maurer [5] reviewed multiple testing methods based on ordered p-values and their mathematical linkage. Recently, Alosh et al. [6] reviewed advanced multiplicity adjustment methods focusing on gatekeeping procedures and graphical approaches in clinical trials. This paper aims to provide an application oriented and comprehensive review of commonly used statistical methodology for multiplicity issues as well as recent developments in multiple testing procedures. Illustrative examples will be used to compare different multiple testing procedures under different scenarios of multiple hypotheses.

This paper will be organized as the following: Section 2 describes notations and two examples to be used in the subsequent sections; Section 3 focuses on non-hierarchical procedures for multiple testing of non-hierarchical hypotheses; Section 4 focuses on multiple testing procedures for hierarchical hypotheses, including methods used for hypotheses with simple ordering (such as the fallback and fixed sequence procedures) and various gatekeeping strategies used for the more complex hierarchical hypotheses; Section 5 reviews the graphical approach which can be used for handling both non-hierarchical and hierarchical hypotheses testing; and Section 6 finishes with a discussion of the merits of each method.

2. Notations and example settings

For the purpose of demonstration, we use the following notations and example settings to illustrate the application of each multiple testing procedure. Let $H_1, \ldots, H_m$ be the null hypotheses, $p_1, \ldots, p_m$ be the raw p-values, and $p_{(1)} \le \ldots \le p_{(m)}$ be the ordered p-values from the smallest to the largest with $H_{(1)}, \ldots, H_{(m)}$ being the corresponding ordered hypotheses. The family-wise error rate (FWER) will be controlled at the α level (two-sided $\alpha = 0.05$). To better illustrate the application of each procedure, we will use two examples of hypothetical trials with two and four hypotheses, respectively. In both examples, two doses are compared with control. The first example considers only one primary endpoint, while the second example considers two endpoints (primary and secondary). Let H_1 and H_2 be the hypotheses on the primary endpoint for Dose 1 and Dose 2, H_3 and H_4 be the hypotheses on the secondary endpoint for Dose 1 and Dose 2, respectively. The first example will be used to illustrate non-hierarchical multiple testing procedures for the non-hierarchical hypotheses in Section 3. Hierarchical testing procedures and the graphical approaches will be illustrated with the second example in Sections 4 and 5. To be specific, we consider scenarios with the following raw p-values:

- Example 1: two dose groups with only one primary endpoint. H_1 is the comparison between Dose 1 and control and H_2 is the comparison between Dose 2 and control. The corresponding raw p-values for the two hypotheses are: $p_1 = 0.04$, $p_2 = 0.024$.
- Example 2 (Fig. 1): two dose groups with one primary endpoint and one secondary endpoint. H_1 is the comparison between Dose 1 and control and H_2 is the comparison between Dose 2 and control for the primary endpoint. H_3 is the comparison between Dose 1 and control and H_4 is the comparison between Dose 2 and control for the secondary endpoint. The corresponding raw p-values are: $p_1 = 0.02$, $p_2 = 0.015$, $p_3 = 0.012$, $p_4 = 0.04$.

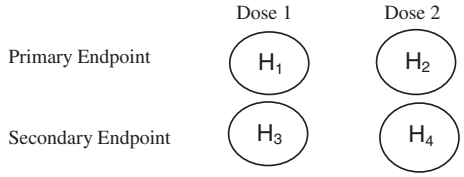

Fig. 1. Illustration of Example 2.

3. Multiple testing procedures for non-hierarchical hypotheses

We categorize multiple testing procedures for non-hierarchical hypotheses as "non-hierarchical" multiple testing procedures which include commonly used procedures such as the Bonferroni procedure, the Holm procedure, the Simes based procedures (Hochberg and Hommel procedures) and the Dunnett procedure. These methods are commonly used in medical research [5]. Descriptions of each procedure and applications with Example 1 follow. For simplicity, we may use the term "accept a hypothesis" to mean "not able to reject a hypothesis".

3.1. Common testing procedures

3.1.1. Non-parametric and semi-parametric procedures

3.1.1.1. Bonferroni procedure. The Bonferroni procedure was first applied by Dunn [7,8] to control the type I error rate for multiple comparisons. It adjusts the significance level for each individual test so that the overall type I error rate is controlled at the desired level. The decision rule for the Bonferroni procedure is: reject H_i if $p_i \le \frac{\alpha}{m}$, for $i = 1, \ldots, m$, where m is the total number of tests. The raw p-values for each test can also be adjusted by multiplying each p-value by m without changing the significance level, which is equivalent to adjusting the significance level for each test. The Bonferroni procedure is widely used due to its simplicity, but it is conservative, especially if m is large and the test statistics are positively correlated.

3.1.1.2. Simes procedure. Simes [9] proposed a modification of the Bonferroni procedure; it rejects the global null hypothesis $H_I = \cap_{i=1}^{m} H_i$ if $p_{(i)} \le \frac{i\alpha}{m}$ for at least one $i = 1, \ldots, m$. The Simes procedure controls the family-wise error rate in the weak sense (i.e., under the intersection of null hypotheses) under independence of the test statistics and the overall type I error rate is controlled if positive regression dependence of the test statistics holds [9,13]. Since the Simes procedure is a global test, it cannot be used for testing individual hypotheses.

3.1.1.3. Holm step-down procedure. The Holm procedure is a non-parametric Bonferroni-based procedure [10]. It is derived via the closed-testing principle. It controls the family-wise error rate without any assumptions on the hypotheses and is uniformly more powerful than the Bonferroni procedure. At the first step, $H_{(1)}$ is tested by comparing $p_{(1)}$ with α/m; if $p_{(1)} > \alpha/m$, then all hypotheses are accepted and testing stops. Otherwise $H_{(1)}$ is rejected and one proceeds to test $H_{(2)}$ by comparing $p_{(2)}$ with $\alpha/(m-1)$. In general, if $p_{(k)} > \alpha/(m-k+1)$, then hypotheses $H_{(k)} \ldots H_{(m)}$ are accepted and testing stops; otherwise, reject $H_{(k)}$ and proceed to test $H_{(k+1)}$. The nominal significance levels for $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ are:

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \ldots, \alpha_m = \frac{\alpha}{1} = \alpha.$$

3.1.1.4. Hochberg step-up procedure. The Hochberg step-up procedure is a semi-parametric Simes-based procedure [11]. It is more powerful than the Holm procedure. It is semi-parametric due to the fact that it only guarantees the control of the type I error rate under independence or certain positive dependence structures of the test statistics [13]. It is slightly more conservative than the Hommel procedure but is widely utilized due to its simplicity for usage in practice. It uses the same nominal significance levels as the Holm procedure but tests the hypotheses in a step-up manner, i.e., it begins with the hypothesis corresponding to the least significant p-value. In general, having accepted $H_{(m)}, \ldots, H_{(k+1)}$, if $p_{(k)} \le \alpha_k$, the procedure rejects $H_{(k)}, \ldots, H_{(1)}$ and stops testing. Otherwise, it accepts $H_{(k)}$ and goes on to test $H_{(k-1)}$.