Chapter 1

Multiplicity Problems in Clinical Trials: A Regulatory Perspective

Mohammad Huque

U.S. Food and Drug Administration

Joachim Röhmel

Bremen Institute for Prevention Research and Social Medicine

1.1 Introduction

Confirmatory controlled clinical trials, also known as Phase III clinical trials, when successful, are significant achievements in medical research as they provide evidence that new treatments (e.g., test drugs or other types of interventions) studied in these trials are clinically effective in treating targeted diseases, and are also safe as far as judgment is possible on taking into account the limited number of exposed patients. Unfortunately, many such trials fail and are unable to show that new treatments studied in these trials are better than placebo. This is surprising because Phase III clinical trials are designed and conducted after the so-called Phase II trials, which are supposed to have already shown promising evidence of treatment efficacy and safety. There can be several reasons for such failures. For example, certain weaknesses in the primary endpoints of a trial can jeopardize the success of a trial, e.g., if these endpoints are not objective, or are not validated, or are not in line with the mechanisms of actions of the treatment. A trial can also fail because of poor planning or disregarding multiplicity issues with respect to multiple endpoints and multiple comparisons.

Clinical trials generally pose multiple questions in the form of hypotheses whose evaluations involve multiple comparisons and tests for multiple endpoints. In testing a single hypothesis, a statistical test in the absence of a treatment effect can lead to a positive conclusion in favor the treatment effect just by chance. Such an error in the testing of hypotheses terminology is

Views expressed in this chapter are the personal views of the authors and not necessarily of any regulatory agency such as the U.S. Food and Drug Administration.

known as a false positive error or a Type I error. When multiple hypotheses are tested without an appropriate adjustment, this error can become excessive. In other words, the familywise error rate (FWER) defined in Section 2.2 can become inflated. This situation can then lead to an easy approval of an ineffective treatment. Therefore, it is important that trials control this error probability at a prespecified level through appropriate design and analyses strategies that are prospectively planned.

This introductory chapter discusses multiplicity problems that commonly arise in confirmatory controlled clinical trials, and also briefly reviews new more efficient methods for dealing with these problems. These methods are described in more detail in subsequent chapters. The types of multiplicity problems that arise in clinical trials depend on how these trials are designed for assessing clinically meaningful benefits of treatments under investigation, given that these trials may collect data on multiple endpoints at different time points on patients randomized to test and control groups at different dose levels. Some trials may also perform multiple interim analyses during the course of the trial and change some design features adaptively based on the results of interim analyses. This creates additional layers of multiplicity problems. Some trials may also be multiregional for which treatment benefit evaluations may require addressing regional differences, e.g., differences in clinical practice, patient ethnicity and other clinical and biological factors. These trials may pose challenging design and multiplicity problems, when in the absence of consistency of treatment effects, evidence of treatment benefits may be sought for specific regions or sub-populations.

With the above background, Section 1.2 introduces the most common multiplicity problems in confirmatory trials. This includes multiplicity problems arising in trials with multiple endpoints when treatment efficacy conclusions are made through the so-called clinical "win" criteria. A win criterion is basically a clinical criterion that defines what sort of endpoint results need to be observed in a trial for a specific disease indication in order to conclude that the treatment provides clinically meaningful treatment benefits. Further, this section reviews multiple comparison problems in trials with a single primary endpoint, e.g., trials with multiple doses compared to a common control or trials with multiple subgroups, and more advanced multiplicity problems, e.g., trials with ordered multiple primary and secondary endpoints. Note that this chapter does not deal with special multiplicity topics, such as multiple comparison problems in trials utilizing adaptive designs. These problems are discussed in Chapter 6.

Here is an outline of the remaining sections. Section 1.3 discusses methods that can be used to reduce the burden of multiplicity in clinical trials, including the use of composite endpoints. Section 1.4 addresses multiplicity concerns for a few special situations that frequently occur in clinical trials such as the use of multiple patient populations. Section 1.5 reviews multiplicity issues for safety endpoints. Finally, Section 1.6 provides some concluding remarks.

Each main section, where applicable, provides necessary motivating ex-

amples and classification of the problems at hand before providing a more detailed overview of relevant multiple comparison problems.

1.2 Common multiplicity problems in clinical trials

This section begins with a review of issues related to the analysis of multiple endpoints in confirmatory clinical trials. The topic of multiple endpoints is discussed in greater detail in Chapter 4. Clinical trials generally classify the endpoints into primary, secondary and exploratory types. Primary endpoints address primary objectives of the trial. They are usually few but are clinically most relevant to the disease and the treatment under study. They assess the main clinical benefits of the treatment. This is usually done through one or more clinical "win" criteria. Examples of such criteria and discussion of multiplicity problems they pose are given in Section 1.2.1. Section 1.2.2 introduces a consistency-based approach to the analysis of multiple endpoints. Section 1.2.3 discusses a specific type of win criterion stating that all primary endpoints individually must show statistically significant treatment benefit. Primary endpoints of this type are usually referred to as co-primary endpoints.

Secondary endpoints characterize extra benefits of the treatment under study after it has been demonstrated that the primary endpoints show clinically meaningful benefits of the treatment. D'Agostino (2000) and Moyé (2003) give detailed classifications of secondary endpoints. O'Neill (1997) supported the idea that "secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance." Section 1.2.4 discusses multiplicity adjustment issues for secondary endpoints.

Exploratory endpoints, for reasons of their own, are usually not prospectively planned and are generally not rigorously evaluated like primary and secondary endpoints. These endpoints are used in treatment comparisons and also unplanned subgroup analyses with an exploratory (e.g., hypothesis generating) purpose. In certain situations, their results can be useful in designing future new trials. However, they are not useful for confirmatory purpose, as there are no valid ways for controlling the FWER for their results. Results of such analyses have the potential of being observed by chance alone.

Further, Sections 1.2.5 to 1.2.8 review other common multiplicity problems in confirmatory clinical trials. Section 1.2.5 discusses multiplicity problems in dose-response and dose-control comparison trials. Section 1.2.6 describes approaches to multiple comparison problems in trials with planned subgroup analyses, e.g., subgroups defined using certain validated biomarker classifiers that are expected to enjoy better treatment efficacy and/or safety. Sections 1.2.7 and 1.2.8 cover multiplicity issues in drug combination and gold standard trials, respectively.

1.2.1 Clinical win criteria with primary endpoints

At times, win criteria are also called "clinical decision rules" for determining clinically meaningful treatment efficacy. They simply define how a positive clinical decision regarding the effectiveness of a test treatment in a trial is going to be reached. The criteria are defined relative to one or more relevant clinical primary endpoints in the setting of comparing one or more doses of test and control treatments. They represent a clinically acceptable way of assessing the effectiveness of a treatment for a disease and patient population under consideration. If statistically significant results are achieved satisfying such criteria, then they can justify meaningful clinical benefits of the study treatment. Although this section focuses on multiple endpoints, the same principles and arguments can be applied to clinical trials with multiple objectives, for example, trials with more that two treatments or trials with inference at more than one time point. Some examples of possible win scenarios with multiple endpoints are as follows:

Example 1. One specified primary endpoint needs to be significant.

Example 2. Given $m \geq 2$ specified primary endpoints, at least one needs to be statistically significant.

Example 3. Given $m \geq 2$ specified primary endpoints, all need to be statistically significant.

Example 4. Given three specified primary endpoints E1, E2 and E3, either E1 needs to be statistically significant or both E2 and E3 need to be statistically significant.

Example 5. Given three specified primary endpoints E1, E2 and E3, either (both E1 and E2) or (both E1 and E3) need to be statistically significant.

Example 6. One of the two specified endpoints must be statistically significant and the other one needs to be at least marginally significant.

Example 7. One of the two specified endpoints must be statistically significant and the other one needs to show non-inferiority.

Example 8. Several endpoints are in a hierarchy that specifies the order in which they are to be evaluated, and significance of all preceding endpoints is required in order to proceed to the next one.

Example 9. Of the m specified primary endpoints, at least k endpoints need to be statistically significant and the remaining $m - k$ endpoints trend in the right direction.

Example 10. There are complex logical relationships among specified primary and secondary endpoints, e.g., secondary null hypotheses need to be tested only if all primary null hypotheses are rejected or at least one primary null hypothesis is rejected.

Examples 1, 3 and 8 do not require any adjustment for multiplicity. For these three examples, each test can be performed at the same significance level of α. Example 1 is the single primary endpoint case, as such, there is no multiplicity issue for this case. Example 3 is the case of co-primary endpoints where all endpoints must show treatment benefit. Testing for treatment benefits for co-primary endpoints follows the intersection-union principle which is discussed later in this section and also in Section 2.3.2. In Example 8, the region of rejection of each test falls within the region of rejection of the previous test whose size is α. Therefore, in this sequential testing scheme (see Section 2.6.3) the FWER always remains bounded above by α. Example 8 type of decision strategy can be useful in getting an earlier onset claim of a treatment benefit in addition to treatment benefits for the usual clinically required onset and durability endpoints. For example, a hair growth trial may test first in sequence the 12th and 15th month endpoints for the usual onset and durability claims of the treatment effects. After this claim is achieved, then one can go back and test the earlier 9th month endpoint for an earlier benefit of the treatment for an additional claim.

Example 2 is the case where each primary endpoint can on its own characterize a clinically meaningful benefit of the treatment under study. Therefore, a win in at least one of the primary endpoints is a win for the trial and the treatment benefits can be labeled as such depending on which endpoints show statistically significant treatment benefits after multiplicity adjustments. Testing for this type of win criterion follows the union-intersection principle which requires multiplicity adjustments for FWER control. This testing principle is covered in Section 2.3.1. For example, a cardiovascular trial may achieve its primary objective for a clinical benefit, if it establishes the efficacy for the study treatment either for all-cause mortality or for MI (myocardial infarction) or for stroke. Thus, the win criterion for this example is to win for efficacy on at least one of the three endpoints. However, in a trial like this when treatment effect sizes are expected to be small but clinically meaningful, an alterative approach is to define a composite endpoint as the primary endpoint. This and the caveats in creating composite endpoints are addressed in Section 1.3.

In Example 4, the null hypothesis is an intersection of two null hypotheses. The first one is a simple null hypothesis that there is no treatment effect for the endpoint E1. The second one is a union null hypothesis that there is no treatment effect for at least one of the two endpoints E2 and E3. The intersection of these two null hypotheses can be tested using the Bonferroni procedure introduced in Section 2.6.1. For example, test E1 at the significance level of α_1 and test E2 and E3 each at the significance level of $\alpha - \alpha_1$. An alternative approach would be to test E1 first at a reduced significance level of α_1 (e.g., $\alpha_1 = 0.025$) and if the result is statistically significant at this level, then test each E2 and E3 at the full significance level of α (e.g., $\alpha = 0.05$), else, test each E2 and E3 at the reduced significance level of $\alpha - \alpha_1$. This latter test is the fallback procedure (see Section 2.6.4). Both approaches will control

FWER in the strong sense, the latter one with a better power for situations when there is a treatment effect for E1.

In Example 5, (E1, E2) and (E1, E3) are pairwise co-primary. This win criterion is equivalent to win for E1 and win for either E2 or E3. In this case, E1 is considered the most relevant endpoint for treatment benefit, but a statistically significant result for E1 alone is considered insufficient to characterize the clinical benefit of the treatment. The clinical decision rule requires an additional statistically significant result in favor of the treatment in at least one of the other two endpoints E2 and E3. In this case, the null hypothesis for the most relevant endpoint E1 and the intersection null hypothesis for the other two softer endpoints E2 and E3 are hierarchically ordered – there is no need for testing for endpoints E2 and E3 unless the null hypothesis for the endpoint E1 is rejected first. Therefore, one can test E1 at the full significance level of α, and if this result is statistically significant then proceed to test for E2 and E3 using a FWER controlling method (e.g., Hochberg or Holm procedures described in Section 2.6) at level α.

Example 6 is the case of two co-primary endpoints, where one endpoint must show convincing evidence of efficacy, and the other one at least marginally. Such a relaxed evidence criterion for a pair of co-primary endpoints is meaningful only if it is clinically acceptable in providing adequate evidence of efficacy for the treatment and the disease under study. Methods for relaxing the evidence criteria to improve the power of co-primary endpoint tests are discussed later in this section and in Section 4.5.

Instead of relaxing the evidence criterion for one endpoint one could set (non-inferiority) margins and require superiority in the more important endpoint and at least non-inferiority for the less important endpoint (Example 7). Because there are two possibilities to win, a multiplicity adjustment needs to be applied to control the FWER (Tamhane and Logan, 2004; Röhmel et al., 2006; Logan and Tamhane, 2008). Superiority/non-inferiority procedures for multiplicity problems of this type are discussed in Section 4.6.

The decision rule in Example 9 has been used in a few arthritis trials and requires appropriate multiplicity adjustments. An arthritis trial generally includes four key endpoints, namely, joint counts for tenderness and swelling, and physician and patient global assessments. As all these endpoints are largely impacted by the condition of the patient's joints, clinical expectations have been that all these four endpoints should show some evidence of treatment benefits. However, because of small treatment effect sizes in some of these endpoints that can result in lack of power of the tests, a clinical decision rule used was to win in at least three out of the four endpoints with one endpoint at least trending in the right direction. However, recent arthritis trials, instead of testing for multiple efficacy endpoints test for a single responder endpoint. This endpoint known as ACR20 combines patient outcomes from seven efficacy endpoints to a single responder endpoint based on a clinical decision rule that determines whether a patient has responded to the treatment

or not. Felson et al. (1995) gave the definition of this endpoint with some validation results.

Example 10 encompasses multiplicity problems of clinical trials that are more complex. They require setting up hybrid multi-stage testing procedures that combine serial and parallel gatekeeping testing strategies and take into account applicable logical restrictions. Problems of this type are discussed later in this section and in Chapter 5.

It should be apparent from the above examples that a null hypothesis for a given clinical decision rule can be complex. A clinical decision rule for winning basically defines an alternative hypothesis space and its complement is then the null hypothesis space. Its statistical testing, when done efficiently with adjustment for multiplicity, can provide clear evidence of clinically relevant benefits of the study treatment.

1.2.2 Consistency-ensured analysis of multiple endpoints

Clinical trials may include more than one clinically important endpoint, each with ability to characterize clinically meaningful benefits of the treatment under study. However, to keep the study size and/or its duration feasible, such a trial may designate only one of these clinically important endpoints as primary leaving others with unknown power properties. Trials of this type can lead to unsettling situations, if the result for the designated primary endpoint is not statistically significant but the result for another alternative clinically important endpoint is strongly positive. Such unsettling situations can arise because of unresolved issues regarding:

- The alternative endpoint, although it produces a strongly positive result, is not a prospectively planned alternative primary with proper α adjustments for evaluating its efficacy.
- The designated primary endpoint result is much weaker than expected or is in the opposite direction causing difficulties in interpreting study findings.

Clinical trial literature reports several trials of this type that had difficulties in the interpretation of study results. An example is the carvedilol trials which were discussed by the FDA Cardiovascular and Renal Drugs Advisory Committee in May 1996. Details about this can be found in Packer et al. (1996) and Fisher and Moyé (1999). In this case, the experimental treatment failed to show a statistically significant result for the planned primary endpoint, namely change in the exercise tolerance, but the results revealed a striking reduction in mortality. Moyé (2003), Moyé and Baraniuk (2007) and Li and Mehrotra (2008) mentioned several other cases of the type where trials were powered for a designated primary endpoint but had interests in testing for another alternative clinically important endpoint as well. The issue of specifying an alternative endpoint can be resolved through a pre-planned

statistical testing strategy that controls the FWER adequately for evaluating both the designated and the alternative endpoint results. Literature has recently introduced several approaches that address this problem, including the prospective alpha allocation scheme (PAAS), nonparametric and parametric fallback methods defined in Chapter 2. The PAAS method has some power advantage over the weighted Bonferroni approach. The nonparametric fallback method for ordered hypotheses gains additional power by allocating a higher α level to a hypothesis if the hypothesis earlier in the sequence is rejected. Similarly, the parametric fallback method attempts to gain additional power over the regular fallback method by taking into account the correlation among the endpoints. However, although the fallback method and its extensions are attractive in saving α for testing additional hypotheses, they do not go beyond the PAAS approach for increasing the chance of a positive trial if a trial does not establish efficacy on the first endpoint. Also, none of these approaches address the issue of interpretation of the study findings when the results for the designated primary and the alternative endpoint are either far apart from each other or are in opposite directions. Recent statistical literature has introduced some methods that weigh the evidence of the two clinically important endpoints of a trial for concluding treatment efficacy. The adaptive alpha allocation approach, called the 4A method by Li and Mehrotra (2008), is an interesting start in that it incorporates the result of the primary endpoint in making inference about the alternative clinically important endpoint. However, further research is needed for better understanding the properties of this method for trials when clinically important endpoints are likely to be correlated. Note that this method mimics the PAAS but the allocated α for testing the second hypothesis is calculated adaptively in a pre-specified manner based on the observed p-value for the first endpoint, so that a weaker result in the primary endpoint requires a much stronger result for the alternative endpoint.

On the other hand, Song and Chi (2007) and Alosh and Huque (2007, 2009) proposed approaches for subgroup analyses that require establishing a certain degree of evidence of efficacy on the primary analysis (i.e., for the total population) before proceeding to test for its subgroup. Huque and Alosh (2009) applied this concept of "consistency" in testing for the designated primary and its alternative endpoint.

1.2.3 Co-primary endpoints

Many trials characterize clinically meaningful efficacy of a new treatment through a single primary efficacy endpoint. However, trials for certain diseases do this through multiple primary efficacy endpoints requiring a statistically significant benefit of the new treatment on each of these endpoints. In clinical trials terminology such primary endpoints are usually referred to as co-primary endpoints. A non-significant result in any of the specified co-primary endpoints would then lead to a non-win scenario for the trial. For example, Alzheimer's trials in mild-to-moderate disease generally include ADAS-Cog and CIBIC

(Clinician's Interview Based Impression of Change) endpoints as co-primary. The ADAS-Cog endpoint measures patients' cognitive functions while the CIBIC endpoint measures patients' deficits of activities of daily living. For a claim of clinically meaningful treatment benefit for this disease, clinicians generally require demonstration of statistically significant treatment benefit on each of these two primary endpoints.

Clinical considerations, such as what benefits a patient in a clinically meaningful way, generally drive the inclusion of two or more relevant primary endpoints in a trial as co-primary. These endpoints are generally used for diseases that manifest in multiple symptoms. All relevant symptoms of the disease must be controlled for a treatment to be viable for such a disease. For example, patients with migraine experience severe headache, typically associated with photophobia, phonophobia, and nausea/vomiting. If a migraine trial shows evidence of treatment benefit only for the endpoint of "headache" and not for other endpoints, then it may earn claim of treatment benefit for headache, but it may fail to do so for the treatment of "migraine." Clinically meaningful treatment benefit claim for migraine, at least in the U.S. regulatory setting, usually requires that the treatment, besides being safe, is effective in relieving all the necessary clinical symptoms of migraine including headache. Additional clinical trial examples of co-primary endpoints can be found, for example, in several CHMP guidelines available at

http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm

and in Offen et al. (2007).

The use of co-primary endpoints in a trial puts extra burden on the trial as it raises the bar for the evidence of efficacy - one must show evidence of efficacy in more than one specified primary endpoints. This generally causes enlarging the size of the trial on carefully assessing that each co-primary endpoint has adequate sensitivity for detecting a desired treatment benefit. Section 4.5 evaluates increases in the sample size in trials with co-primary endpoints in relation to that for single primary endpoint trials. However, despite this sample size concern, many co-primary endpoint trials are successfully done. A reason for this is that the endpoint treatment effect sizes for these trials are fairly reasonable, often 0.3 standard deviations or greater (see Huque et al., 2008). In addition, some of these trials being symptomatic trials are not difficult to conduct with respect to patient enrollment and endpoint ascertainments. Table 1.1 gives some idea of the sample sizes for trials with 2 to 4 co-primary endpoints in relation to single endpoint trials, when effect sizes for co-primary endpoints are in the range 0.2 to 0.4 standard deviation. Note that this table is different than that given in Section 4.5 in that it gives some idea about the size of the trial in terms of the endpoint effect sizes. This table shows that the co-primary endpoint trials can be feasible with respect to sample sizes if the endpoint treatment effect sizes are about 0.25 standard deviation or better. Note that trials with other types of win criteria that require multiplicity ad-

TABLE 1.1: Sample sizes per treatment arm in a trial with m co-primary endpoints ($\alpha = 0.025$, one-sided tests, power = 80%). The test statistics follow a multivariate normal distribution with equal pairwise correlations and common effect size.

Effect size	Correlation	Single endpoint trial	Trial with m co-primary endpoints		
			$m = 2$	$m = 3$	$m = 4$
0.20	0.0	393	516	586	636
	0.4		497	556	597
	0.8		458	494	518
0.25	0.0	252	330	375	407
	0.4		318	356	382
	0.8		293	316	332
0.30	0.0	175	230	261	283
	0.4		221	247	266
	0.8		204	220	230
0.40	0.0	99	129	147	159
	0.4		124	139	150
	0.8		115	124	130

justments also decrease power and require sample size increase in comparison to trials with a single endpoint.

Statistically, the null hypothesis for testing co-primary endpoints is a union null hypothesis and the alternative an intersection hypothesis. A test of such a null hypothesis is sometimes referred to in the statistical literature as an intersection-union test (see Section 2.3.2). This test does not inflate the Type I error rate, and as such, there is no downward adjustment of α for this case, i.e., each co-primary endpoint test can be performed at the same significance level of α. If each of the m co-primary endpoints is tested at the significance level of α then the maximum Type I error rate for testing m co-primary endpoints is the same as α. However, testing for co-primary endpoints inflates the probability of the Type II error, and thus, reduces the power of the test.

In testing for co-primary endpoints, the inflation of the Type II error rate, and consequently the reduction in the power of the test, besides depending on m, also depends on the extent of dependency between the test statistics of the co-primary endpoints. If these test statistics were independent, the power would simply be the product of the powers for single endpoint tests, with Type II error probability as one minus this power. If the test statistics were perfectly correlated, and the treatment effects sizes (per unit standard deviation) for the co-primary endpoints were homogeneous, then there would be no inflation of the Type II error probability. As the correlation between test-statistics can fall in between the two extremes, so can be the power of the co-primary endpoint tests. Therefore, in designing clinical trials with co-primary endpoints, one must also consider the dependency between the test statistics for addressing the Type II error probability.

In testing for co-primary endpoints, treatment effect sizes also impact the power of the test. In a trial, one may not expect the same treatment effect size for all co-primary endpoints. Some endpoints, such as low event endpoints, can have relatively small effect sizes in comparison to other endpoints, causing the trial to be powered with respect to the low yield endpoints. However, in some such situations, this difficulty can be overcome to some extent by adequately enriching the trial, or alternatively, low yield endpoints, if they are likely to exhibit similar treatment effects, by combining them to a composite endpoint, thus reducing the dimensionality in the co-primary endpoint testing. Such a composite endpoint can be acceptable if it is clinically relevant for the disease under study and has regulatory and scientific acceptance. Section 1.6 addresses considerations for composite endpoints.

An important consideration in designing a clinical trial with co-primary endpoints is to be sure that the compound under study is such that it has the ability through its mechanism of actions to target all the co-primary endpoints of the trial. Before launching such a trial, scientists usually validate this assertion through animal studies, early phase human trials (such as proof-of-concept trials) and on synthesizing historical trials of similar compounds. A trial with co-primary endpoints will surely fail if an endpoint in the co-primary set has no efficacy sensitivity for the given compound.

1.2.4 Secondary endpoints

As indicated at the beginning of this section, the role of secondary endpoints in confirmatory clinical trials is considered different than that of the primary endpoints. Primary endpoints address primary objectives of the trial. On the other hand, secondary endpoints have a number of important functions at levels different than those of primary endpoints. Details of these functions of secondary endpoints can be found in D'Agostino (2000) and in Moyé (2003). It is well understood that secondary endpoint hypotheses can be tested for extended treatment benefits after the primary objectives of the trial have been successfully met. However, at times, for planning purposes, a key endpoint such as mortality is called a secondary endpoint on expecting it to be a low yield endpoint. In this case, such a secondary endpoint is like a primary endpoint and is sometimes called a key secondary endpoint. Its result, after proper multiplicity adjustments, if statistically significant in favor of the treatment, can provide a persuasive evidence of a clinical benefit of the study treatment.

Different clinical trials depending on the objectives often take different approaches towards multiplicity adjustments for the analysis of secondary endpoints. These approaches vary from no adjustments to adjustments with strong control of the FWER. If the purpose of secondary endpoint analyses is to make additional claims of treatment benefits, in addition to those already established by the analyses of the primary endpoints, then multiplicity adjustments generally require a strong control of the FWER. In this regard, a useful analysis approach is the gatekeeping approach. On the other hand,

if the purpose of secondary endpoint analyses is simply to facilitate interpretations of the results of the primary endpoints, and there is no intent for additional claims of treatment benefits, then a strong control of the FWER for secondary endpoint analyses may not be necessary. However, regardless of the approach taken for the analysis of secondary endpoints in a specific trial, there is a general agreement that the secondary endpoints along with their methods of analysis should be prospectively planned, and the outcomes of these endpoints should be carefully ascertained similar to those for primary endpoints. Following are some of the approaches that seem to have emerged in the context of the analysis of secondary endpoints.

Approach 1. Control the FWER in the strong sense at a pre-specified significance level α in testing both the primary and secondary endpoint families of hypotheses. Test these two families of hypotheses hierarchically, and in this regard, consider the family of primary endpoint hypotheses ahead of the family of secondary endpoint hypotheses. This can allow endpoint specific claims of treatment benefits for secondary endpoints after the primary objectives of the trial have been met. In this case, the multiplicity adjustments for the secondary endpoints can depend on the results of the primary endpoints, but normally the multiplicity adjustments for the primary endpoints should not depend on the results of the secondary endpoint tests.

Approach 2. Once the primary objectives of the trial are met, then test for the family of secondary endpoints independently of the tests for the primary endpoints, but adjust for multiplicity by controlling the FWER for the family of secondary endpoints in the strong sense. In this case, the multiplicity adjustments for the secondary endpoints do not depend on the results of the primary endpoints – all that is needed is that the primary objectives of the trial have been met. This approach can result in an inflation of the FWER under Approach 1.

Approach 3. Analyze secondary endpoints only for supporting evidence without any intention for a claim of treatment benefit for the secondary endpoints. In this case, the p-values, confidence intervals, and any other results for secondary endpoints are purely for descriptive purposes and not for drawing inferences.

1.2.5 Dose-response and dose-control comparisons

Dose-response studies in clinical settings are essential for finding a dose or dose range of a treatment which is efficacious as well as safe. However, the methods and objectives of the dose-finding analysis differ for Phase II versus Phase III trials. In Phase II trials dose-finding is generally model-based. In Phase III trials, on the other hand, dose-finding is generally based on hy-

pothesis testing, requiring pre-specification of a suitable multiple comparison procedure given the trial design and the nature of claims planned for the trial.

In Phase II trials, one fits a set of suitable parametric dose response models, such as $E_{\max}$ or logistic models and their variants, given dose-response data on a key response variable at m doses of the test treatment including placebo as a zero dose. Fitting of these models requires estimation of the parameters of the models. A working model, which is the best candidate for describing the dose response relationship, is then selected from these fitted models by performing appropriate statistical tests. This working model is then used for proof of activity of the test treatment in the dose range of the trial and for deciding which few doses should be included in the confirmatory trial. In this approach, the need for multiplicity adjustments arises for addressing uncertainty in the model selection. Bretz et al. (2005) described this approach in detail. This approach has the advantage that it incorporates clinical pharmacology model-building concepts in setting candidate models and inferences for dose selection are not confined to the target doses among the dose levels under investigation. See Section 3.5 for more information on methods that combine multiple comparison procedures with modeling techniques.

An alternative conventional approach is based on hypothesis testing. This approach treats dose as a qualitative factor with very few assumptions, if any, about the underlying dose response model. This approach is often used for identifying the minimum effective dose (MED) that is statistically significantly superior to placebo and is clinically relevant (see Tamhane et al., 1996; ICH E4, 1994). Target dose estimation is discussed in Section 3.3. This approach, though relatively robust, does not take into account the clinical pharmacology concepts that drive the selection of suitable dose response functions to be considered for a specific treatment and disease situation. The use of model-based approaches is emphasized for Phase II trials for deciding about a therapeutic dose or dose range for confirmatory trial use.

In a Phase III setting, a general design for multiple doses is that which includes two or more doses of the test treatment, placebo, and one or more doses of active control. Bauer et al. (1998) investigated this design in detail and have proposed various inferential strategies for dealing with questions of interest and multiplicity issues. Other designs are mainly special cases of this general design as follows:

Case 1. Trials that include 2 or more doses of a test treatment and a placebo, without any active control arm, with (a) no order restriction among doses and with (b) order restriction among doses. When there is no order restriction among doses, the global test, also called the overall heterogeneity test, can be performed, for example, by one-way analysis of variance with a common unknown variance under the normality set up, contrast-based trend tests (see Section 3.2) or by the Kruskal-Wallis test under the non-parametric set up. This global test can be replaced by the standard Dunnett procedure or Dunnett-Tamhane (1991) step-down procedure for comparing different doses to placebo (see Section

2.7). However, when order restriction among doses can be safely assumed (e.g., dose effects are in non-decreasing order by dose), one can use the Bartholomew or Williams tests discussed in Section 3.2. Alternatively, one can perform a fixed-sequence procedure presented in Section 2.6.3.

Case 2. Trials with two or more doses of the test treatment and one or more doses of an active control. These trials are unable to include placebo because of ethical reasons. Such a trial design is used for many serious diseases like meningitis infection or serious cardiovascular diseases when placebo treatment can cause irreversible harm to patients. For these trials, often it is either well-known to the disease area experts or is evident from the relevant historical data that the active control treatment efficacy benefit is substantially larger than that for placebo. For this case, the efficacy goals can then be accomplished without the placebo in the trial by the statistical non-inferiority testing methods.

Case 3. Trials with three or more doses of the treatment without placebo and without active control. This type of trial is used for establishing efficacy of a new treatment when it is unethical to include placebo in the trial and also an appropriate active control is unavailable for setting the trial as a non-inferiority trial. For such a trial, an approach for establishing efficacy of a treatment is to establish its dose response, e.g., one of the higher doses of the treatment is superior to its lower dose in efficacy. The underlying assumption is that the highest dose may not be sufficiently tolerable to have adequate compliance for demonstrating efficacy. However, at least one of the higher doses is safe, tolerable and efficacious, but some lower doses may not have sufficient efficacy. Sample size requirement can be large for such a trial. One can test for the global hypothesis for an overall assessment of efficacy followed by pairwise comparisons of dose groups with proper multiplicity adjustments.

1.2.6 Subgroup analyses

Subgroup analyses are quite prevalent and are considered necessary in clinical research. Commonly, the purpose is to either justify consistency of results across clinically relevant subgroups or discover large treatment differences among certain subgroups. Regulatory guidance recommends subgroup analyses by race, gender and age for pivotal comparative trials. Subgroup analyses, however, pose analytical challenges and are generally fraught with difficult inferential issues including multiplicity and lack of power issues. Lagakos (2006) wrote, "Subgroup analyses are an important part of the analysis of comparative trials. However, they are commonly overinterpreted and can lead to further research that is misguided, or worse, to suboptimal patient care." Yusuf et al. (1991), Wang et al. (2007) and others raised similar concerns about subgroups analyses and included some helpful recommendations.

In this section, we discuss multiple comparison aspects of subgroups analysis, and emphasize that prospective planning is key to any subgroup analysis.

Often subgroup analyses are data-driven and the biological plausibility of a positive finding is argued after the data has been analyzed and results seen. Such subgroup analyses produce results that are usually false positives and not replicable. The ISIS-2 trial (ISIS-2 collaborative group, 1988) result discussions include an interesting example in this regard. Unplanned subgroup analyses may not be helpful even for hypotheses generating purposes; it can lead to misguided future research. In this situation, it is extremely difficult to know, or to be convinced of, even if somebody keeps a record of it, as to how many analyses were done prior to finding a significant positive result. Even ignoring the issue of bias for questionable comparisons, it is extremely arduous to control the Type I error rate for such analyses without knowing the correct number of analyses performed and the pre-specified adjustment method. If the investigator was honest, and, say, he did 20 analyses in searching for a significant p-value of 0.05 or less, then the Type I error for this result could be as high as 64 percent. Another key point often ignored is that in unplanned subgroup analyses, treatment groups may not be comparable because of randomization issues causing confounding and bias in the results. Consequently, p-values for treatment comparisons cannot be validly interpreted.

Frequently subgroups formed are "improper subgroups." An improper subgroup is defined as a group of patients characterized by a variable measured after randomization and potentially affected by treatment and study procedures. For example, an improper subgroup analysis in clinical trials is the so called "completers analysis" that excludes treatment failures. In contrast, a "proper subgroup" is a group of patients characterized by a common set of "baseline" characteristics that cannot be affected by treatment, e.g., patient demographic characteristics or prognostic disease characteristics defined before randomization.

Moyé (2003) and others argued that if multiple subgroups are analyzed, then the best estimates of the treatment effects in subgroups are not the treatment effects observed in those subgroups; rather they should be shrunk to the average treatment effect for the total patient population of the trial; see Efron and Morris (1977). Low or high treatment effect sizes in subgroups can appear even under the null hypothesis of no treatment-by-subgroup interaction because of the regression to the mean phenomenon. On the other hand, a targeted subgroup within a trial may have an atypical treatment effect. This information may be available from prior clinical trials or from Phase II trials, or from biological reasoning, e.g., from histology types, as to why a targeted subgroup may have potential to show better treatment efficacy than the rest of the patients of the trial. For example, a positive outcome of a pharmacogenomic biomarker for a patient may identify that patient at randomization to be a potential responder to the given treatment.

A subgroup analysis, whether planned or unplanned, can experience a significant loss of power in detecting a treatment effect. However, this loss may

not occur for a targeted subgroup whose effect size happens to be sufficiently greater that that for the total patient population of the trial. By using a power formula due to Koch (1997), Alosh and Huque (2008) showed that for some situations when the treatment effect size for a subgroup exceeds that of the total population, the power for the treatment effect test (after proper adjustment for multiplicity) for that subgroup can exceed that for the total population treatment effect. This observation, and the previous work by Simon and Maitournam (2004) and Maitournam and Simon (2005), opens the door for designing trials for testing for treatment effects for a planned targeted subgroup for which one expects a much larger size treatment effect than that for the total patient population. Traditional randomized trials are usually not well-planned for subgroup analyses. Such an approach may be adequate if the study treatment effect is expected to be homogeneous across subgroups. However, if this treatment effect is expected to be heterogeneous, such trials may miss identifying subgroups of patients which are most likely to benefit by the study treatment.

Composite nature of Phase III trials with respect to patient characteristics is well-recognized. Consequently, the extent of treatment efficacy of an intervention can be different in different subgroups of patients. For example, patients with non-fatal MI and stroke are likely to respond better to a treatment if treated early than late. A genetic mapping and testing of virus in a virus-infected patient can tell whether this virus type will be resistant or susceptible to a given treatment. A breast cancer with estrogen receptor positive outcome can respond better to a treatment than a breast cancer estrogen receptor negative outcome. Herceptin responds better for metastatic breast cancer patients with an HER2 protein over-expression (Burstein, 2005). Thus, a trial, during randomization, can be enriched by a subgroup of patients who are likely to respond better to a given treatment than the rest of the patients of the trial. This can increase the success of the trial and can make the test more powerful for testing the treatment effect for the targeted subgroup. However, in this case, there is often concern that the overall treatment effect is driven mainly by the treatment effect in the subgroup and there is no treatment effect for the complement subgroup. If the review of the data of the complementary subgroup suggests that this is a possibility then the product label may reflect this. For example, the product label for atenolol (Tenormin) based on subgroup analyses included the statement, "Some subgroups (e.g., elderly patients with systolic blood pressure below 120 mm Hg) seemed less likely to benefit" (see Physicians' Desk Reference, 2002, Page 693). Such complimentary efficacy subgroup concern can also be addressed on using consistency-ensured multiple testing strategies discussed in Section 1.2.2.

1.2.7 Combination drug trials

For fixed-combination drug products regulatory guidance requires that each component product, as mono-therapy, in the combination must be ef-

ficacious and safe and the combination product must demonstrate superior efficacy to each of its components in order to justify clinically meaningful benefit of the combination. This has led to trial designs with three arms, one arm for the combination and the other two for each of the components, provided that each component is an approved product for efficacy and safety. However, if the efficacy of any of the two components can not be assumed in a trial (e.g., for symptomatic treatments) or any of the two components is an unapproved product, then a 2-by-2 factorial design, which includes a placebo, is sometimes used. In this case, there is an extra burden to show that the unapproved component has clinically acceptable safety profile in addition to showing that it has a clinically meaningful efficacy.

The statistical test for these designs is the usual intersection-union test for showing that the combination is statistically significantly superior to each component, and if placebo is required in the trial, each component is also statistically significantly superior to placebo. Each of these tests is performed at the same significance level (Laska and Meisner, 1989; Sarkar et al., 1995).

The win criterion for efficacy for a drug combination trial can become complicated if two or more primary endpoints or three or more components are required in a trial to show the benefit of the combination over its approved components (Kwak et al., 2007). For example, for the case of two endpoints, the win criterion for efficacy of the combination may be to show superiority for one endpoint and at least non-inferiority for the other endpoint when comparing the combination to each component (superiority/non-inferiority procedures for problems of this kind are discussed in Section 4.6).

Sometimes multi-dose factorial designs are employed for the assessment of combination drugs for serving dual purpose, to provide confirmatory evidence that the combination is more effective than either component and to identify an effective and safe dose combination or a range of useful dose combinations (Hung, 2000).

1.2.8 Three-arm trials with an active control

In many therapeutic areas, well-established standard treatments exist and for all new treatments seeking market authorization for the same indication there are then choices for the control treatment in clinical trials. If a standard treatment (active control) is selected this would lead to a two-arm study comparing the new treatment to the standard treatment for non-inferiority (using an appropriate non-inferiority margin) and (possibly) for superiority. There are well-known weaknesses connected with this design (see ICH E9) and therefore its use is recommended only for diseases where it is unethical to include placebo in the trial, and if sufficient historical data are available for the active control that can help in resolving the issues that these trials pose. These weaknesses come from the complexities of non-inferiority trial designs.

The goal for a standard non-inferiority trial is to infer indirectly about the efficacy of the new treatment on demonstrating that the new treatment is close

or similar (within a certain margin of non-inferiority) to the active control, which itself has previously been demonstrated to be effective by being superior to placebo. Thus, these trials have two comparisons, a direct comparison of the treatment against the active control and an indirect comparison against placebo which the trial is not able to include. However, the validity of the direct comparison, depends upon the validity of the indirect comparison and how much is known and how much can be assumed about the treatment effect of the active control in the setting of the current trial. Therefore, for assuring validity of this type of comparisons, the non-inferiority trials are required to have the following properties:

- Assay sensitivity of the trial. It is the ability of the trial to have shown a treatment difference of the active control in comparison to placebo of a specified size, if the trial had a third arm with placebo control. Without this property, the trial has the undesirable ability to conclude that the treatment is non-inferior to an ineffective drug.

- Constancy assumption. The trial is sufficiently similar to past historical studies with respect to all design and conduct features that can influence the estimation of the treatment effect of the active control, e.g., in regard to design features, patient population characteristics, important concomitant treatments, definition and ascertainments of study endpoints, dose of active control and background regimen, entry criteria, and analytic methods.

- Quality of the trial. This is in the interest of ruling out undesirable conduct features of the trial that would tend to minimize the difference between the treatment and the active control causing bias toward the null hypothesis. These include, for example, imprecise or poorly implemented entry criteria, poor compliance and the use of concomitant treatments whose effects may overlap with the treatment, in addition, inadequate measurement techniques, or errors in treatment assignments

Therefore, because of the above difficulties associated with non-inferiority trial designs, the use of three-arm trials with the new treatment, active control and placebo has gained much wider attention for diseases where it is ethical to include placebo in the trial. This type of trial has been called the "gold standard" design. The CHMP guideline "Choice of the non-inferiority margin" (2005) recommends this design to be used wherever possible. Pursuing all aims simultaneously in this type of trials raises the issue of multiple testing. There are several comparisons of interest:

- New treatment versus placebo for superiority.

- Active control versus placebo for superiority.

- New treatment versus active control for non-inferiority and (possibly) for superiority.

This approach is discussed in Koch and Röhmel (2004) and Röhmel and Pigeot (2009). See also Pigeot et al. (2003), Hauschke and Pigeot (2005) and discussion papers to this article. According to the closed testing methodology one could begin with an overall test for any differences between the groups which are followed by separate individual comparisons using the same significance level as for the overall test. Alternatively one could start with a test tailored to the many-to-one-situation (new treatment and active control versus placebo).

Pigeot at al. (2003) started with the comparison of the active control with placebo. Only after the active control has been shown to exhibit the expected superiority over placebo, Fieller's theorem (1954) is applied to the ratio of "net differences" (new treatment minus placebo and active control minus placebo). The limits of the resulting confidence interval allow a precise quantification of the fraction f of the effect of the active control that is preserved by the new treatment. A fraction $f > 0$ is translated into "new treatment is superior to placebo", a fraction close to 1 means similar effects of the active control and new treatment, and a fraction larger than 1 confirms superiority of the new treatment over the active control.

If one is willing to accept a hierarchical structure between the hypotheses, one could also proceed as follows. Because the performance of the new treatment is of primary interest one could start immediately with the comparison between the new treatment and placebo. If the new treatment failed to demonstrate superiority over placebo, the remaining comparisons would lose much of their importance and would mainly serve for interpretation. Any superiority of the active control over placebo would not be surprising because it did so consistently in the past, and would reinforce the view that the new treatment is sufficiently effective. A failure of the active control, however, would raise doubts in the study itself (conduct, patient selection, sample size, or other design features).

The test for non-inferiority of the new treatment versus the active control is meaningful only in the case that one can successfully establish the superiority of the new treatment over placebo. Then it follows logically that two further null hypotheses of interest (the new treatment is inferior to the active control and the active control is no more effective than placebo) cannot both be true simultaneously. This means that both hypotheses can be tested using the full significance level, independently of each other. If the new treatment is non-inferior to the active control, a final test for superiority of the new treatment versus the active control can be carried out, again at the full level α. This procedure controls the FWER in the strong sense. A similar problem was considered by D'Agostino and Heeren (1991) for which Dunnett and Tamhane (1992) showed how the sequence of the tests should be conducted so that the FWER is controlled. Active controlled trials that do not include placebo for ethical reasons, usually test for non-inferiority first, and if non-inferiority is established, then test for superiority next. This fixed-sequence testing has both clinical and statistical merit. The reverse fixed-sequence test-

ing, that is, test for superiority first and then for non-inferiority second, is generally not recommended for this type of trials. A reason for this is that the primary method for such a trial is usually the indirect demonstration of efficacy of the new treatment, as indicated above. In this framework, the non-inferiority hypothesis is primary and the superiority hypothesis is secondary. The sample size and design considerations of trials with this focus, and also the trial conduct rules and interpretation of results, are quite different than superiority trials. In addition, the determination of assay sensitivity and the non-inferiority margin can be quite complex and challenging.

1.2.9 Advanced multiplicity problems

Multiple comparison problems discussed in the preceding sections dealt with a single source of multiplicity, e.g., two-arm trials with multiple endpoints or single-endpoint trials with multiple doses or subgroup analyses. However, multiplicity problems become more complex in trials with multiple sources of multiplicity, e.g., multiple treatment arms, multiple endpoints and tests for non-inferiority and superiority. The purpose of this section is to give examples of some such complex multiple comparison problems for some of which satisfactory solutions for regulatory applications have either not been worked out or available workable solutions are not sufficiently clear.

Example 1. A trial compares two doses of a new treatment to a control with respect to two primary efficacy endpoints and a quality of life (QoL) endpoint with all three endpoints considered equally important. The QoL endpoint is a composite (global) endpoint which includes four components that serve as secondary endpoints and can be tested only if the null hypothesis for the QoL endpoint is rejected.

Example 2. A trial compares four dose levels D1, D2, D3 and D4 of a new treatment to placebo on two primary endpoints for finding which dose levels are significantly better than placebo for both endpoints. In this comparison, the higher dose levels D3 and D4 are primary dose levels of equal importance and lower dose levels D1 and D2 are included to better understand the dose-response relationship in case the two higher dose levels are significant on both endpoints.

Example 3. A trial tests for treatment effects for the onset and durability endpoints at low, medium and high doses of a new treatment compared to placebo. In addition, the trial tests for certain loss in efficacy of effective doses at durability endpoints for putting patients on some sort of maintenance dose for continued meaningful treatment efficacy.

Example 4. A trial tests for treatment effects for multiple primary and secondary endpoints at low, medium and high doses of a new treatment compared to placebo with the restriction that tests for the secondary

endpoints for a specific dose can be carried out only when certain primary endpoints show meaningful treatment efficacy for that dose.

Example 5. A trial compares a new treatment to an active control on a primary and a secondary endpoint, and in this comparison, the trial performs non-inferiority and superiority tests in succession for the primary endpoint first. If non-inferiority or superiority are established for this primary endpoint, the trial then performs similar tests for the secondary endpoint.

Example 6. Instead of a single primary and a single secondary endpoint, as in Example 5, the trial has multiple primary and multiple secondary endpoints, and the non-inferiority or superiority test are carried out for certain secondary endpoints only if all primary endpoints show non-inferiority or superiority of the treatment against this control.

Example 7. A trial with a single primary endpoint includes three doses (high, medium and low) and an active control with non-inferiority and superiority tests for these doses. In this comparison, the trial sponsor faces two situations:

- Case 1. Dose effects can be assumed to be of non-decreasing order with increasing dose.
- Case 2. The assumption of a non-decreasing dose effect cannot be made, except perhaps in some special cases where only the low dose can be assumed to have efficacy not exceeding those of the medium and high doses.

A number of methods have been proposed to address the multiplicity problems in Examples 1 through 7. Hommel et al. (2007) proposed procedures based on the Bonferroni test that address multiplicity problems in Examples 1, 2 and 7. Further, procedures based on more powerful tests, e.g., Hochberg- or Dunnett-type tests, can be constructed using the general tree gatekeeping framework described in Section 5.5 (see also Dmitrienko, Tamhane and Liu, 2008). This framework allows testing of hierarchically ordered multiple families of hypotheses with logical restrictions. However, further research is needed to work out the details. Also note that solutions to the above problems can be far more challenging if trials were to also include interim analyses and allowed changes in certain design features based on interim results.

As an illustration, consider the multiple testing problem in Example 7. In Case 1, one may proceed with a fixed-sequence test, i.e., test for the high dose first for non-inferiority and then for superiority, and proceed to test similarly for the medium dose and then for the low dose if each time the preceding dose is found to be at least non-inferior to the active control. It has been argued that in this sequential testing, as two tests are performed for each dose, FWER can exceed α if each test is performed at the same level α. Therefore, for strong

control of FWER at level α, appropriate multiplicity adjustment is warranted for this problem. Further, in Case 2, several approaches are possible. One approach would be to test first for non-inferiority simultaneously for all three doses, and then test for superiority only for those doses found to be non-inferior. A second approach would be to test for non-inferiority and then for superiority separately for each dose at adjusted α levels, e.g., at $\alpha/3$ for the Bonferroni method, or at levels that account for correlations among the test statistics. A third approach would be to define a family F_1 of tests for the high and medium doses only and a family F_2 of tests for the low dose only. Then assign α_1 to F_1, where $\alpha_1 < \alpha$ (e.g., $\alpha_1 = 2\alpha/3$), and perform tests for non-inferiority and superiority for each dose in F_1 as performed in the second approach above (e.g., spend $\alpha/3$ for each dose test in F_1). If a non-inferiority or superiority is established for a dose in F_1 then α used for that is basically saved which at least in part can then be carried forward to F_2 on satisfying strong FWER control at level α. As an example, if x doses in F_1 were found to be superior to the control with the Bonferroni adjustment of $\alpha/3$ for each dose, where $x = 0$, 1 and 2, then the low dose in F_1 can be tested at level $\alpha(x+1)/3$.

Example 8. A clinical trial requires multiplicity adjustments for comparing two doses of a new treatment to a placebo on two primary endpoints. These endpoints are able to characterize a clinically meaningful benefit of the treatment for each of these two doses with the condition that if only one of the endpoints is statistically significant for a dose then the other endpoint should exhibit at least a weak evidence of treatment benefit, that is, its p-value if greater than 0.05 is less than some pre-specified number α^* (e.g., $\alpha^* = 0.25$). In addition, for this situation (i.e., when only one of the endpoints is statistically significant), a much stronger evidence of treatment benefit is required for the other endpoint. This problem can be solved by proper modifications of the consistency-ensured methods proposed in Section 1.2.2.

Example 9. A clinical trial uses a surrogate endpoint S for an accelerated approval and a clinically important endpoint T for a full approval. It uses α_1 for the accelerated approval and $\alpha_2 = \alpha - \alpha_1$ for the full approval. In this case, a larger allocation for α_1 will increase the chance of an accelerated approval; however, it would impact negatively on the final approval under the underlying assumption that S is predictive of T only when the expected treatment effect for S exceeds certain limit. For example, an expected treatment effect, say of size x, for S would establish efficacy for S at a pre-specified level $\alpha_1 = 0.01$. However, such a value of x for S may not predict appropriate efficacy, say of size y, for the final endpoint T at $\alpha_2 = 0.04$. An appropriate procedure needs to be developed to take into account the strength of evidence in S (e.g., in terms of its p-value and other information) that will increase the chance

of a positive trial for the final endpoint T and will control the Type I error rate.

1.3 Reducing multiplicity in clinical trials

Testing many hypotheses in a trial can easily overload a trial with an excessive multiplicity burden making the trial unnecessarily complex, large and unrealistic from the cost and public health perspectives. Therefore, for meeting the objectives of a trial, consideration should be given in its design to reduce the burden of multiplicity efficiently. In this regard, several approaches are possible. A popular approach, discussed in Section 1.3.1, is to hierarchically order the families of hypotheses and also, as far as possible, the hypotheses within each family. This allows the use of new innovative statistical methods that adjust for multiplicity much more efficiently, in some cases with minimal or no adjustments at all. In addition, the use of dependence information among the test statistics can help in reducing the extent of multiplicity adjustments. This follows from the fact that the inflation of the FWER is generally largest when the test statistics are statistically independent, but relatively small when they are highly correlated. Another approach introduced in Section 1.3.2 is to combine several endpoints to define a single or a few composite endpoints.

1.3.1 Hierarchical testing

Families of null hypotheses are said to be hierarchically ordered or ranked if earlier families serve as gatekeepers in the sense that one tests hypotheses in a given family if the preceding gatekeepers have been successfully passed. The two commonly used hierarchical families of endpoints in a clinical trial are the family of primary endpoints and the family of secondary endpoints. These two families are hierarchically ordered with the property that rejections or non-rejections of null hypotheses of secondary endpoints depend on the outcomes of test results of primary endpoints. The individual endpoints within a family can also have hierarchical ordering, occurring naturally or by design. Hierarchical ordering of multiple endpoints and also of multiple comparisons can considerably reduce the multiplicity burden in controlling the FWER in a trial. The following examples further illustrate these points.

Example: Gastric ulcer trial

Consider a gastric ulcer healing trial that compares a new treatment to a control using two endpoints that are healing of gastric ulcers (verified by endoscopy) and the symptom of gastric pain at the end of 16 weeks of treat-

ment. These two endpoints are naturally hierarchically ordered. That is, if the treatment heals all gastric ulcers in a patient then there is likely to be complete amelioration of gastric pain as well for that patient. In this case, one can use the fixed-sequence procedure (Section 2.6.3), i.e., test the healing endpoint first at the significance level of α, and if this test is statistically significant then test for the symptom endpoint at the same significance level of α. Both endpoints being statistically significant can strengthen the efficacy of the treatment of the trial.

Example: Congestive heart failure trial

Consider a congestive heart failure (CHF) trial that accumulates events of mortality and hospitalizations over a two-year treatment period and has two primary endpoints. One endpoint is the composite of mortality and hospitalizations and the other is the mortality-only endpoint. These two endpoints can be ordered by design to test for the composite endpoint first thinking that it will have more events over the two-year treatment period for better power of the test for this endpoint. This situation, as in the above example, allows testing of the two endpoints sequentially at the same significance level.

However, a danger in sequential testing is that it can miss an important result. For example, if testing stops after the first test (i.e., the two-sided p-value for the composite endpoint is greater than 0.05) and the 2-sided p-value for the mortality endpoint happens to be small (e.g., less that 0.01) then this will incur a loss of a valuable trial result. For discussions following the results of a large clinical trial with outcomes for the two endpoints similar to this one, see

http://www.fda.gov/ohrms/dockets/ac/03/briefing/
3920B2_02_A-FDA-Coreg.pdf

This danger can be reduced by using the fallback procedure discussed in Section 2.6.4. The fallback procedure has the flexibility of moving forward in the testing sequence even if some or all of the preceding tests in the testing sequence happen to be statistically non-significant. For example, for such a trial, one could test the composite endpoint at a pre-specified significance level of 0.035, and if the composite endpoint result is statistically significant at this level, then test the mortality-only endpoint at the full significance level of 0.05; otherwise, test the mortality endpoint at the reduced significance level of 0.015.

Example: A dose-finding hypertension trial

For a rather complex example of hierarchical families of hypotheses involving multiple endpoints and multiple dose groups, consider a dose-finding study in patients with hypertension. Such a study was conducted to evaluate the effects of three doses, low, medium and high, of a treatment compared to

placebo where the effects were measured by reduction in systolic and diastolic blood pressure (SBP and DBP) measurements. For this study it was reasoned that

- SBP is more indicative of the true effect than DBP and hence it was placed at higher hierarchy.
- The medium and high doses were considered equally important, and potentially equally relevant, while the lower dose was considered less likely to exhibit significance.

Therefore, hierarchically ordered families of null hypotheses for this trial were set up as follows:

- $\{H_{11}, H_{12}\}$: Null hypotheses for testing each of the high and medium doses compared to placebo with respect to SBP.
- $\{H_{21}, H_{22}\}$: Null hypotheses for testing each of the high and medium doses compared to placebo with respect to DBP.
- $\{H_{31}\}$: Null hypothesis for testing the low dose compared to placebo with respect to SBP.
- $\{H_{32}\}$: Null hypothesis for testing the low dose compared to placebo with respect to DBP.

A parallel gatekeeping technique in Dmitrienko, Offen and Westfall (2003) provides a solution for this complex hierarchical testing problem (gatekeeping procedures are discussed in Chapter 5); an alternative solution based on the parametric fallback procedure is given in Huque and Alosh (2008) (see Section 4.3).

1.3.2 Composite endpoints

Another popular approach for reducing multiplicity, particularly for cardiovascular trials, is to combine several clinically relevant endpoints into a single endpoint, known as a composite endpoint. Chi (2005) cited three types of composite endpoints for clinical trials as primary endpoints:

- An index or a responder endpoint is constructed from multiple item scores, counts or from low yield endpoints. The HAM-D total score for depression trials and the ACR20 endpoint for rheumatoid arthritis trials are of this type. However, such a composite endpoint has to be clinically valid for the disease and the type of intervention under study. In addition, there should be sufficient experience from historical clinical studies to show that it has been reliable in detecting treatment effects, and has also been interpretable and clinically meaningful.

- A failure rate after a certain period of treatment or follow-up is computed on counting events that can occur from a predefined set of events. For example, in an organ transplant trial a failure during the six month of treatment can be a biopsy-proven acute rejection or graft loss or death.

- A few binary events are combined to form a composite event endpoint. Such a composite event endpoint generally arises in cardiovascular and chronic disease trials.

Chi (2005) made a distinction between the second and third types, but we see them as of the same type. Components of a composite endpoint are usually referred to as component or singleton endpoints. If E1 and E2 are binary events, then a composite event E is the union of these two events; i.e., it counts as an event if either E1 or E2 alone occurs, or both E1 and E2 jointly occur. For convenience in counting, one often counts only the first occurrence of the component events in a patient.

There are several motivations for using a composite event endpoint as a primary endpoint. It can reduce the size of the trial if the following conditions are met. The components in the composite increase the number of events in a non-overlapping manner; i.e., an event is not a direct consequence of the other. In addition, there is evidence for some homogeneity of treatment effects across the components of the composite or the components jointly enhance the treatment effect. A composite endpoint can also address broader aspects of a multifaceted disease. In this case, an isolated result in an endpoint may not be clinically meaningful. A composite endpoint can also allow changing the focus of the trial from discovering a large treatment effect to clinically meaningful small treatment effects that collectively can achieve a statistically significant benefit of the treatment. For further informative discussions of these motivations and relevant examples, see Moyé (2003).

Development of a composite endpoint at the time of trial design, however, requires special considerations and prior empirical evidence of its clinical value. Components of the composite are supposed to have some consistency with regard to their importance to patients, frequency of their occurrence, and the extent of treatment effects they are likely to produce. The use of a composite endpoint is generally discouraged for confirmatory clinical studies when large variations are expected to exist among its components with regard to these three aspects. The results of the component endpoints must also be fully disclosed and displayed along with the results of their composite for allowing a meaningful interpretation of the composite endpoint result.

Each component of a composite endpoint conveys a special meaning to the clinician. Therefore, statistical analysis of a composite endpoint generally includes some sort of analysis for its components. This raises multiplicity issues. However, statistical testing in this case usually follows a hierarchical structure; i.e., one tests for the composite endpoint first and then for its components. However, the multiplicity adjustment strategy for the components, depend-

ing on the objectives of the trial, can vary. It can vary from no adjustment for the analysis of components to adjustments to control the FWER in the strong sense. Therefore, multiple approaches for the analysis of components exist depending on the purpose of such analysis. Following are a few testing strategies that one can adopt in this regard.

Testing Strategy 1. In this case, the composite endpoint is tested at a prespecified significance level α and component endpoints are not statistically tested but their results are summarized for the interpretation of the composite endpoint using descriptive statistics that generally include point, interval estimates and unadjusted p-values for treatment effects. This approach has the disadvantage that if one of the components has a sufficiently small p-value, no specific claim of benefit can be made for this specific component endpoint, as there was no prospective plan for multiplicity adjustments for such a claim.

Testing Strategy 2. Use a sequential testing scheme. That is, if the composite endpoint result is statistically significant at the significance level of α, then test for specified components in sequence at the same significance level α. This method controls the FWER in the strong sense in testing all components. However, no claim can be made for a component even if it has a very small p-value if the sequence breaks, that is, when a test that precedes the testing of this particular component is not rejected. This deficiency can be avoided by using the fallback testing strategy, which allows testing of an important component of a composite, such as death, even when the null hypothesis for the composite endpoint is not rejected (see Section 2.6.4).

Testing Strategy 3. This is similar to Testing Strategy 2, except that instead of using a sequential testing scheme, one uses another method that controls the FWER in the strong sense. Examples include multiple testing procedures described in Section 2.6.

In resolving treatment effects, some components of a composite may have more persuasive abilities than others. The ones that have less persuasive abilities are often referred to as "soft components" of a composite endpoint. Showing a treatment benefit for the death endpoint (hard endpoint) can be more persuasive than showing a treatment benefit for the hospitalization endpoint. If the overall treatment effect is supported predominantly by the soft components then there should be assurance that the hard components are not influenced negatively. The following three approaches can address the issue of soft components of a composite endpoint.

The sub-composite approach. Create a "sub-composite" of those components of the "main composite" that are hard components. Then there are two null hypotheses to test one for the main composite and the other for the sub-composite. One can then test the null hypothesis for the main

composite at a smaller significance level α_1 and the sub-composite at a larger significance level α_2, such that $\alpha_1 + \alpha_2 = \alpha$. For the above example, one may select $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$.

Non-inferiority method. Define a priori an acceptable margin for "sub-composite" and test for superiority for the main composite and test for non-inferiority for the "sub-composite", both at the same significance level.

The weighting method. In this method, computation of the treatment effect estimate and the test statistic uses pre-determined α "weights" given to the component endpoints with the sum of the weights being equal to one. For example, the weights selected for the harder endpoints could be three times larger than the weights assigned to softer endpoints. The advantage of this method is that it avoids adjusting for an extra sub-composite endpoint. However, these weights must be pre-defined and acceptable to clinicians.

There is a widespread interest for using composite endpoints as primary endpoints in a trial either for a single or multiple comparisons, especially for the cardiovascular type trials and for trials with patient reported outcomes that target claim of treatment benefits for a specific or multiple domains of a disease. This is because of the attractive nature of the composite endpoints in reducing multiplicity problems and in reducing the sample size of the trial. However, a cautionary point worth emphasizing in this regard is that trials with composite endpoints can be complicated when components of such endpoints are of widely differing importance to patients and occur with differing frequency, and are likely to vary markedly in the magnitude of treatment effects. Such situations can complicate interpreting the results of a trial. Therefore, when large such variations are likely to exist across components of composite endpoints, their use for controlled clinical trials can be problematic for making specific claims of treatment benefits.

1.4 Multiplicity concerns in special situations

1.4.1 Use of multiple subsets of patient data

It has been a common practice for many years that the primary statistical analysis concerns all randomized patients. The ideal situation that all patients are treated as intended by the randomization schedule and that all patients adhere to the study protocol is rare. Some irregularities usually occur. Patients may drop out of the study early, or patients may receive treatments different from the one assigned to them by the randomization schedule. Questions then arise as to how to deal with these irregularities. Because the true

measurements are not available, one must make assumptions. Often more than one plausible assumption is possible and sometimes a conservative view may require making more pessimistic assumptions. Therefore, several analyses must be performed on the same endpoint but with varying subsets of patients and/or varying subsets of data. As the reason of such "sensitivity analyses" is to demonstrate that the main study results are only marginally influenced by the observed irregularities, no adjustment for the Type I error rate is necessary. However, a different situation arises when following a statistical analysis for the whole study population subgroups are selected and subjected to separate statistical analyses. Multiplicity issues arising in this situation are discussed in detail in Section 1.2.6.

1.4.2 Use of multiple statistical methods

Different statistical models or statistical methods are sometimes applied to the same data. In a parametric regression model one could, for example, include or exclude explanatory variables, or one could apply transformations to the data for achieving symmetry. It is also possible to change the analysis strategy from a parametric model to a nonparametric one or vice versa. Usually the primary analysis strategy is laid down in the Statistical Analysis Plan (SAP) that is finalized before the randomization schedule is opened. The analysis strategy as laid down in the SAP should contain sufficient details to avoid any ambiguity with the primary analysis. Technically one could also split the Type I error rate, if different analysis strategies are applied with control of the FWER. However, this has been rarely applied. In contrast, sometimes several analysis strategies are applied with the aim to demonstrate that the results are robust and are present regardless of the statistical model or method. Usually such procedures are not preplanned and as long as they do not indicate discrepant results they at least do no harm.

1.4.3 Analysis of covariance

Unplanned analysis of covariance, similar to unplanned subgroup analysis (see Section 1.2.6), is prone to producing misleading results because of inherent multiplicity, bias and other issues. For this reason unplanned analysis of covariance is generally discouraged for drawing confirmatory evidence from a trial. Nonetheless, it is considered useful for exploratory purposes and for interpreting the results of a primary analysis. It can facilitate clarifying the degree to which the estimated treatment effects are due to the study treatment or by other factors that are associated with the response variable. It can also help in evaluating the consistency of treatment differences across subgroups. On the other hand, planned analysis of covariance is a useful method for randomized controlled clinical trials. If appropriately used, it can serve important purposes. For example, when comparing treatment groups, it can improve the power of the test under certain conditions through variance re-

duction resulting in smaller p-values and narrower confidence intervals. It can also adjust some of the random imbalances in baseline prognostic variables to "equivalence" in comparing treatment groups.

Planning and pre-specification of covariates and other aspects of this analysis, such as model specification, are considered essential. In addition, clear statistical as well as clinical justifications, stated in advance, are required for assuring that such an analysis will provide an unbiased estimate of the true difference between the treatments, and the covariate-adjusted treatment effect can be meaningfully interpreted for a clinical benefit. Covariates pre-selected for such an analysis need to be strongly correlated with the response variable. Covariates measured after randomization may be affected by the treatment – their inclusion into the analysis is likely to produce biased results. The CPMP (2003) document "Points to consider on adjustment for baseline covariates" provides a comprehensive list of recommendations to be followed when planning an analysis of covariance for a confirmatory trial. Also Pocock et al. (2002) and Senn (1994) provided useful comments and recommendations on this matter.

1.5 Multiplicity in the analysis of safety endpoints

This section includes some basic concepts and general framework for analysis of multiple safety endpoints. Safety evaluation of a treatment or an intervention intended for the treatment of a disease is an important objective of all clinical trials in general. For this reason, clinical trials collect adverse events (AEs) data for each patient of the trial along with other relevant information that help in clinical evaluation as well as quantitative analysis of these events. An AE refers to an "untoward medical event associated with the use of a drug in humans, whether or not it is considered as drug-related." Some of these AEs can be serious. They are often referred to as serious adverse events (SAEs) or serious adverse drug experiences (SADEs). Examples of these are death, a life threatening event, an in-patient hospitalization, a persistent or significant disability, a congenital anomaly or a birth defect. Depending on the medical judgment, it may also include interventions such as medical or surgical procedures used for preventing serious safety events.

Statistical analysis of safety data includes analysis of particular events, analysis of event rates, evaluation of risk over time, exploration of possible subgroup differences, and the identification of risk factors associated with serious events. These analyses rely on those quantitative methods that can validly quantify risk and provide an adequate measure of uncertainty. The analysis approach usually taken is to analyze the safety data for each clinical study and then integrate the safety findings across multiple studies and other clinical experiences. In this effort, investigators group closely related events.

This is accomplished by using the so called the "dictionary" of preferred terms, such as the Medical Dictionary for Regulatory Activities (MedDRA), which was developed under the auspices of ICH. These analyses of safety data play an important role in determining the risk-benefit profile of a new treatment.

It is well recognized that analyses of safety data entail multidimensional problems involving many safety endpoints. A key point, however, is that not all these endpoints can be pre-specified. A number of them are unanticipated. This then calls for a different statistical inferential approach for different parts of the safety data. In this regard, it is worth recognizing that design and analysis approaches for safety assessments are in general different than those for efficacy assessments. For efficacy, primary and secondary efficacy endpoints are specified in advance. Trial sizes are usually determined with respect to primary endpoints, and serious efforts are made, in planning interim looks at data or in addressing multiplicity of tests for preserving the Type I error rate. On the other hand, in general, clinical trial practices have been not to specify safety hypotheses or the level of sensitivity in advance. However, there are exceptions. These are for situations when a particular safety concern related to a specific drug or drug class has risen or when there is a specific safety advantage being studied. For these cases, sometimes, trials for efficacy evaluation are large enough to address these safety issues, but every so often, when there is a significant safety concern, special large safety studies are conducted with pre-stated hypotheses and uncertainty specifications similar to efficacy trials. Considering these challenges and the ICH E9 ideas for safety data analysis, one usually classifies trials' safety analysis strategies into three types:

Type 1. Analyses of adverse experiences associated with specific hypotheses that are formally tested in the clinical study and both Type I and Type II error concerns are properly addressed.

Type 2. Analyses of common adverse experiences which are usually treatment emergent signs and symptoms (TESS) not present at baseline, or if present at baseline, a greater severity seen when on treatment. Such AEs usually do not have pre-stated hypotheses, but their causality is assessed on comparing the incidence rates for patients treated with the test treatment versus those treated with the control. If the incidence rate of an AE is clearly greater with the test treatment than with the control, it can be attributed to the test treatment. For this purpose, clinical trials report p-values, point and/or interval estimates of risk differences or relative risks as descriptive statistics and for flagging purposes. However, methods differ when trials are short-term versus long-term studies. Long-term studies usually account for person-exposure times on treatment in the analysis with adjustments for censored observations. The inference approach may differ depending whether the intent is to report non-inferiority versus superiority.

Type 3. These refer to analyses of less common and rare spontaneous reports

of adverse events, some serious, often requiring attention of specialty area experts. Analysis of these usually require large data bases because these events may occur at very low rates, e.g., in the range of 1/100 to 1/1000.

Analysis methods for Type 2 AEs, for addressing multiplicity, are required to have certain special properties. For example, it should provide a proper balance between no adjustment and too much adjustment in the sense that one is willing to tolerate some wrongly flagged AEs provided their number is small relative to the total number of flagged AEs. In addition, it should address Type II error concerns and provide adjusted p-values for flagging purposes. In this regard, Mehrotra and Heyse (2004) recommended methods that provide control of the false discovery rate.

An approach that addresses Type II error concerns is to specify an error rate for failing to identify at least k (e.g., $k = 1$ or 2) unwanted safety events out of the total K analyzed; then for a specified α, work out the size of the safety database needed for this purpose. If this size is limited then adjust the α upward accordingly.

1.6 Concluding remarks

Clinical trials generally include multiple objectives, e.g., multiple efficacy and safety endpoints or comparisons of multiple doses of a new treatment to a control. The objectives are formulated to seek answers to a set of specified scientific questions. Whether answers to these questions can lead to claims of clinically meaningful benefits of the new treatment is determined by multiple "win" criteria that introduce multiplicity. Confirmatory trials prospectively plan statistical procedures that deal with these multiplicity issues in an efficient manner. Regulatory agencies in this regard usually ask for a well-developed and thought-out statistical analysis plan (SAP). If the multiplicity issues are complex, the SAP may include trial simulation results and sometimes mathematical proofs for assuring the validity and efficiency of the proposed statistical procedure.

A key point in this regard is that solutions to multiplicity problems generally require adhering to the principle of prospective planning, i.e., defining the multiplicity problems and working out their solutions in advance. Another key point is that, in the presence of multiplicity of tests, when specific claims of treatment benefits are intended for a new treatment, strong control of the FWER is almost always required.

In conclusion, as the subject matter of multiplicity for clinical trials is vast, and all cannot possibly be captured in an introductory chapter like this one even at the conceptual level, the scope of this chapter has been intentionally

limited to standard multiplicity topics that commonly arise in confirmatory clinical trials. Other multiplicity topics of complexity that arise, for example, for interim analyses and also for adaptive designs of confirmatory clinical trials, are also important multiplicity topics. These have been omitted here, but the reader may find some of them in other chapters of this book.