

Chapter 2

Concept of Alpha

2.1 Lady Tasting Tea

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Alpha Type I Error Rate

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

Table 2.1 Null hypothesis and decision

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

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

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

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

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

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

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

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

2.3 Intention-to-Treat

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

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

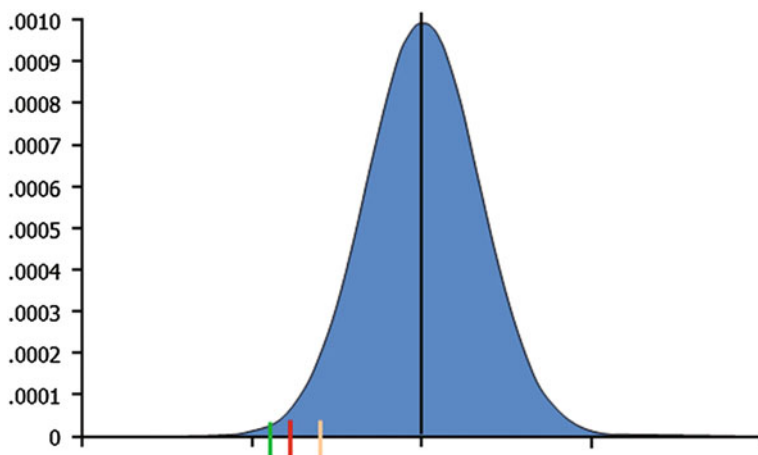


Fig. 2.1 A normal density curve with the critical value on the *left*

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

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

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

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

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

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

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

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

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

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

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

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

2.4 Patient Analysis Sets

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

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

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

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

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

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

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

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

2.5 Multiple Comparisons

2.5.1 Multiple Doses

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.5.2 *Multiple Endpoints*

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

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

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

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

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

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

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

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

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

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

- Bonferroni Method

- $p(i) > \alpha/K$, then “accept” null hypothesis of no effect ($\alpha = \text{alpha}$).
- $p(i) \leq \alpha/K$, then reject null hypothesis of no effect.

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

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

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

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

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

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

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

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

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

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

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

2.5.2.1 Summary Measures or Summary Statistics

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

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

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

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

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

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

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

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

2.5.3 *Other Types of Multiplicity*

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

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

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

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

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

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

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

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

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

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

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

- μ is the overall mean
- τ_i is the i th treatment effect
- γ_j is the j th stratum effect
- β is the slope associated with the baseline measure

X_{ijk} is the baseline corresponding to the k th patient in i th treatment of k th stratum,
and
 ε_{ijk} is the residual of this subject

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

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

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

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

2.6 P -Value and Statistical Significance

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

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

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

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

Reconsider the model equation

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

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

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

Table 2.4 ANOVA

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

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

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

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

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

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

2.7 Stages of a Clinical Trial

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

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

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

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

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

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

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

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

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

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

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

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

2.8 Subject Selection and Choice of Alpha at Phase II

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

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

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

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

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

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

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

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

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

References

- Bellamy, N., Campbell, J., Stevens, J., Pilch, L., Stewart, C., & Mahmood, Z. (1997). Validation study of a computerized version of the Western Ontario and McMaster Universities VA 3.0 osteoarthritis index. *The Journal of Rheumatology*, 24, 2413–2415.
- Cappelleri, J. C., Zou, K. H., Bushmakina, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. P. (2013). *Patient-reported outcomes: Measurement, implementation and interpretation* (p. 201). Boca Raton, Florida: Chapman & Hall/CRC Press.
- Chuang-Stein, C., & Kirby, S. (2014) The shrinking or disappearing observed treatment effect. *Pharmaceutical Statistics*, 13, 277–280.
- Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., & Offen, W. (2005). *Analysis of clinical trials using SAS: A practical guide*. Cary, NC: SAS Institute Inc.
- Dmitrienko, A., Tamhane, A. C., & Bretz, F. (eds.) (2010). *Multiple testing problems in pharmaceutical statistics*. Boca Raton, FL: Chapman & Hall/CRC.
- Dunnnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50, 1096–1121.
- Evans, S., & Ting, N. (2015). *Fundamental concepts for new clinical trialist*. Boca Raton: Taylor and Francis/CRC Press.
- Friedman, L. M., Furberg, C. D., & DeMets, D. L. (2010). *Fundamentals of clinical trials* (4th ed.). New York, NY: Springer.
- Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2(3), 100–112.
- International Conference on Harmonization. (ICH). (1998). E-9 statistical principles for clinical trials.
- Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y., & Mancuso, J. P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal*, 42, 303–319.
- O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40, 1079–1087.
- Piantadosi, S. (2005). *Clinical trials: A methodologic perspective* (2nd ed.). Hoboken, NJ: Wiley.
- Proschan, M., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York: Springer.
- Rothman, K. J., Lash, T. L., & Greenland, S. (2008). *Modern epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Salsburg, D. S. (2002). *Lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Henry Holt and Co.
- US Department of Health and Human Services. (2009). *Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims*. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf>

Phase II Clinical Development of New Drugs

Ting, N.; Chen, D.-G.; Ho, S.; Cappelleri, J.

2017, XVII, 241 p. 25 illus., 17 illus. in color., Hardcover

ISBN: 978-981-10-4192-1