Guideline on Multiplicity Issues in Clinical Trials

(Trial Version)

English Translation by: Bo Yan, Yong Wang, Ping Yin

Disclaimer: The English is for information only and not an official translation and under any dispute the Chinese will prevail

Center for Drug Evaluation, NMPA

November, 2021

Contents

Guideline on Multiplicity Issues in Clinical Trials

1. Overview

Multiplicity issue is common in clinical trials, which refers to the situation where the study conclusion relies on a set of statistical inferences simultaneously (multiple testing). For example, when a study has multiple endpoints (e.g. primary endpoints and key secondary endpoints), multi-group comparisons, multi-stage decision-making (e.g. interim analysis for making efficacy decision), multiple time-point analyses of longitudinal data, subgroup analyses, analyses of different combinations of parameters or analyses on different data sets under the same statistical model, sensitivity analyses, and so on. For confirmatory clinical trials, the basic statistical principle is to control the familywise error rate (FWER) to a reasonable level. It should be noted that some of the above multiplicity issues can inflate the FWER, while others may not. For the former, the FWER needs to be controlled at a reasonable level by using appropriate strategies and statistical methods, the process called multiplicity adjustment; for the latter, no multiplicity adjustment is needed. Therefore, when developing a clinical trial protocol and writing a statistical analysis plan, it is always important to use appropriate strategies and statistical methods to control the FWER.

This guideline mainly describes the common multiplicity issues and their corresponding decision-making strategies, introduces multiplicity adjustment methods and multiplicity analysis methods, aiming to provide guidance on how to control FWER in confirmatory drug clinical trials. The general principles discussed below can also apply to other types of clinical studies.

2. Type I error, familywise error rate (FWER), and type II error in multiple testings

2.1 Type I error and FWER

Type I error refers to the error in which the null hypothesis is true but the test result rejects it, in the clinical trials it is an error drawing the effectiveness conclusion from statistical inference for an ineffective drug. The probability of making such an error should be controlled at a certain level, which is called test level or significance level, denoted by alpha (α). The test level for a single hypothesis test in a multiple hypothesis testing procedure is known as the nominal test level, or the local test level, denoted by α_i.

Considering a clinical trial in which its conclusion needs to be supported by a series of hypothesis tests, the FWER is the probability that at least one true null hypothesis is rejected among multiple hypothesis tests. Regardless of which subsets of null hypothesis are true, the so-defined FWER is controlled, it is called the strongly controlled FWER. When FWER is controlled under the condition that all null hypotheses are true, it is called the weakly controlled FWER. The weakly controlled FWER can only draw overall conclusions but does not support the conclusions regarding single hypothesis testing, so its application in confirmatory clinical trials is not very meaningful. This guideline is only focused on the strongly controlled FWER defined above.

2.2 Type II error

Type II error refers to the error in which the null hypothesis is incorrect, but the test result fails to reject the null hypothesis, in the clinical trials, it is the error of statistically drawing an ineffective conclusion on an actually effective drug. This probability is denoted by the symbol beta (β). Accordingly, 1-β is called testing power. For confirmatory clinical trials, the risk of type II error should also be taken into consideration, provided

that type I error is effectively controlled. For multiple testing, the requirement of controlling FWER reduces the significant level α_i for each individual test, which in turn reduces the testing power. Therefore, when multiplicity adjustment is performed, the development of a study plan should take into consideration the impact of controlling the FWER on the statistical power, for example, by appropriately increasing the sample size to guarantee sufficient power.

3. Common Multiplicity Issues

The common multiplicity issues in clinical trials are generally associated with multiple endpoints, comparisons between multiple groups, subgroup analyses, interim analyses, and analyses of longitudinal data at different time points.

3.1 Multiple endpoints

3.1.1 Primary endpoint

A primary endpoint is an endpoint that is directly related to the primary concern (primary objective) of the clinical trial and that provides the most clinically meaningful and convincing evidence and is commonly used for the main analysis, the sample size estimation, and the evaluation of whether or not the trial has met its primary objective. In confirmatory clinical trials, a single primary endpoint is common, but multiple primary endpoints may be involved in some cases. For studies with multiple primary endpoints, there are generally two types of research hypotheses i.e., each of the multiple primary endpoints is required to be statistically significant or at least one of the multiple primary endpoints is statistically significant.

(1) All primary endpoints are required to be significant. That is, the study drug is considered effective when all results of primary endpoints are significant (often referred to as co-primary endpoints). For example, in a

confirmatory clinical trial for the treatment of chronic obstructive pulmonary disease, there are two separate primary efficacy endpoints, forced expiratory volume in first second and patient-reported symptom score. Both results of primary endpoints must be statistically significant before the study drug can be claimed to be effective. In this case, there is no inflation of FWER because this strategy does not have any opportunity to select one or more of the primary endpoints that are most favorable to the new drug, and there is only one possibility to conclude that the drug is effective (i.e., both null hypotheses are rejected). However, this increases the type II error and hence decreases the testing power. The degree of power reduction is related to the number of primary endpoints as well as the correlation between the primary endpoints. The more the numbers of the primary endpoints and the weaker their correlations, the less the power becomes.

(2) At least one of the multiple primary endpoints is required to be statistically significant. In this circumstance, the study drug is considered to be effective as long as at least one of the primary endpoints is statistically significant. For example, in a confirmatory clinical trial designed to show the effectiveness of a drug for the treatment of burn wounds, two separate primary endpoints are used: wound closure rate and scar formation. The clinical trial protocol specifies that the drug is considered clinically effective as long as at least the result of one of the endpoints is statistically significant. There is an inflation of FWER in this case because the conclusion that the drug is effective includes three possible scenarios: a significant rate of wound closure without significant scar formation; a non-significant rate of wound closure with significant scar formation; and both are significant on the rate of wound closure and scar formation. Because the conclusion of the effectiveness of the new drug may be declared due to various combinations of significant primary endpoints, whether or not it

will lead to the inflation of FWER depends on the specific research hypotheses.

3.1.2 Secondary endpoints

Clinical trials often have multiple secondary endpoints, which in most cases intend to provide support for the primary endpoint. On some occasions, the secondary endpoints may be used to support the benefits claimed in the labeling, and are generally referred to as key secondary endpoints. In this case, the key secondary and the primary endpoint should be considered together in control FWER. Hypothesis testing of the key secondary endpoints will be performed only if the hypothesis testing of the primary endpoint is considered globally significant.

3.1.3 Composite endpoint

The composite endpoint refers to the combination of multiple clinically relevant outcomes into a single variable, where the endpoint is defined as the occurrence in a patient of any one of the specified components. e.g. considering a composite endpoint relating to cardiovascular events: myocardial infarction, heart failure, and sudden coronary death where the occurrence of each event implies the occurrence of the composite endpoint; or the scores of several symptoms and signs will be combined into a single variable through certain methods, such as the ACR20 scale for the evaluation of rheumatoid arthritis trials. If a composite endpoint is used as a single primary endpoint, multiplicity will not be an issue. However, if a component of the composite endpoint (e.g., an event or a dimension of a questionnaire) is also used to support the benefit claimed in the labeling, it should be treated as a primary or key secondary endpoint and the multiplicity adjustment should be taken into consideration.

3.1.4 Exploratory endpoints

Exploratory endpoints may be prespecified or non-prespecified (e.g., some data-driven). It often includes clinically important events that are expected

to occur at such a low frequency that a treatment effect is difficult to be shown, or endpoints that are considered unlikely to show the effects for other reasons but are included for exploring hypotheses, the results of which may help design new clinical trials in the future. Multiplicity issues would not happen in these scenarios.

3.1.5 Safety endpoints

If a safety endpoint (event) is a part of a trial's confirmatory strategy to support the benefits claimed in the labeling, it should be specified in advance and considered for the multiplicity issue. It should be noted that in clinical trial practice, due to the large uncertainty of safety events, it is sometimes difficult to specify the primary safety hypothesis in advance. Therefore, the confirmatory strategy for controlling multiple safety endpoints (usually serious adverse reactions) may be based on the post hoc multiplicity adjustment strategy, which should be fully justified and a consensus should be reached with the regulatory authorities.

3.2 Comparison among multiple groups

Comparisons between multiple groups are common in clinical studies, examples include three-arm designs, dose-response relationship studies, and studies for combination therapies and fixed combination drugs, and so on.

3.2.1 Three-arm design

The three-arm design is often seen in non-inferiority trials, where the three arms are assigned to the test group, positive control group, and placebo group. Three scenarios should be considered for the research hypothesis: ① superiority of the comparison between the test group and the placebo, ② superiority of the comparison between the positive control group and the placebo, and ③ non-inferiority of the comparison between the test group and the positive control group. For the multiplicity concern, if the

investigational drug can be considered effective relies on the rejection of all three hypotheses, or as long as it meets the requirement of ① (this strategy needs to be agreed upon by regulatory authorities before it can be implemented), or a fixed sequential strategy is adopted, such as sequence hypothesis testing ①→②→③, then it will have no FWER inflation. For other three-arm designs which do not follow this multiplicity test strategy and do not require all test results are significant, it is necessary to consider whether it will lead to an inflation of FWER.

3.2.2 Dose-response relationship

Dose-response relationship studies are essential to finding safe and effective therapeutic doses or dose ranges. The methods and objectives of dose-finding analysis are different between the exploratory trial and the confirmations trial.

In the exploratory trials, when conducting the dose-finding studies, it is up to the sponsor to decide whether or not to control the FWER. In the confirmatory clinical trials, to select and confirm one or more doses recommended for the investigational product in a specific patient population, the FWER must be controlled.

3.2.3 Combination therapies and fixed combination drugs

Combination therapy refers to the use of two or more therapeutic drugs at the same time for the treatment. A fixed combination drug refers to a drug with a combination of two or more compounds for the treatment. The aim of a clinical trial of combination therapies or fixed combination drugs is primarily to verify whether the benefit-risk profile of the combination is better than that of the individual component.

Take the combination therapy of two single drugs as an example, at least three treatment groups will be included in the trial design, namely combination therapy group, single drug A group, and single drug B group,

and the latter two groups are served as the positive control groups. If an additional placebo group is added, it is a 2 × 2 factorial design. No Matter using a three-group design or a four-group factorial design, statistical tests to infer whether the combination therapy is superior to the other groups will not lead to the inflation of FWER because the efficacy of the combination therapy will be demonstrated only when all null hypotheses are rejected.

3.3 Analysis of longitudinal data at different time points

Longitudinal data, i.e., data collected from repeated measurements at different time points, is common in clinical trials. Analyses of such data in relation to time points are performed in two cases, one is comparing the group treatment effect across different time points; the other is comparing treatment effects across different time points within a group.

For example, a study design that has only one primary endpoint and involves only two treatment groups, if the evaluation of the primary endpoint between treatment groups is only performed at one of the multiple time points (e.g., the last visit), and the comparison between groups at other time points is considered as the evaluation of the secondary endpoints, no multiplicity issue is involved. If the evaluation of the primary endpoint between treatment groups is performed at more than one time point, and the differences are required to reach significance at all relevant time points, then there is no FWER inflation, but if the differences not reach the above significance, then there will lead to the inflation of FWER. In the case of comparing effects at different time points within a treatment group, a multiplicity issue needs to be considered if the objective is to confirm the effect at the optimal time point by comparison between time points, i.e., when the time effect becomes part of the confirmatory strategy.

For study designs with more than one primary endpoint or more than two treatment groups and involving analysis of longitudinal data at different

time points, the multiplicity problem is more complex and requires comprehensive consideration.

If one wishes to avoid the problem of multiplicity issue for longitudinal data, one possible solution is to transform the effects at different time points into an area under the curve, for example, pain VAS scores at different time points after treatment can be transformed into areas under the curve to represent the total pain score after treatment, in another word, multiple variables are transformed into one variable. However, after the transformation, between-group comparisons at each time point may not be implemented. Another possible solution is to analyze repeated measure data with a single model, such as repeated measure analysis of variance or mixed-effects models.

3.4 Subgroup analysis

Subgroup analyses are generally conducted to show the efficacy of the drug in a target subgroup population or the consistency of efficacy results across subgroups. If the purpose of the analysis of a target subgroup is for supporting the benefits claimed in the labeling, the multiplicity issue raised from the total and the subgroup population analysis needs to be considered together and one needs to ensure enough subjects in the subgroup so that the subgroup analysis has sufficient power. If subgroup analyses are not used to support the benefits claimed in the labeling, no multiplicity issue needs to be considered.

3.5 Interim analysis

When conducting the interim analysis for checking the efficacy, because multiple decisions need to be made in the research process and the multiplicity issues are complex and diverse, it is particularly important to control FWER. Appropriate FWER control strategies and corresponding methods should be carefully considered and prespecified when developing clinical trial protocols.

3.6 Complex design

Complex designed confirmatory studies such as basket designs, umbrella designs etc., and platform designs cover multiple disease areas and multiple drugs across different disciplines. The multiplicity issue is involved due to the simultaneous development of multiple sub-topic studies. However, since these sub-topic studies are mostly independent and answer specific clinical questions, such as applicable diseases and target population, it is generally not lead to the inflation of FWER.

However, when there is a large overlap in the target population of complex design sub-studies, or when the same control group is used for multiple sub-studies, whether it leading to the inflation of FWER should be determined according to the specific situation, and adequate communication between the sponsor and regulatory authorities is recommended.

4. Common strategies and methods for multiplicity adjustment

For the multiple issues that may lead to FWER inflation in clinical trials, the strategy and method of multiplicity adjustment depend on the study objective, the design, the hypothesis, and the test method used in the trial. The sponsor shall make a necessary evaluation of the selected strategy and method for multiplicity adjustment in the trial design and make a detailed description in the clinical trial protocol as well as the statistical analysis plan.

The choice of strategies and methods for multiplicity adjustment can be considered from three aspects: decision-making strategy, adjustment method, and analysis method.

4.1 Decision strategies for multiplicity issues

The conclusions of clinical trials are mainly inferred based on the results of the analysis of all trial data. It is also a process from local decision-

making to overall decision-making. The decision strategies for multiplicity issues can be divided into parallel strategies and sequential strategies. In addition to the process of moving from local to global decision-making, there are staged global decisions. First, the possible multiplicity issues can be sorted out according to the study objective and study design. Then one strategy or a combination of different strategies can be adopted, and finally, the statistical analysis methods used for each hypothesis test and the corresponding allocation strategy of nominal test level α_i can be determined according to the selected strategy or combination of strategies (if necessary).

4.1.1 Parallel strategy

Parallel strategy means that the included hypothesis tests are independent of each other and are performed in parallel, independent of the testing order. Like a parallel relationship, the results of each hypothesis test do not depend on the results of other hypothesis tests.

4.1.2 Sequential strategy

Sequential strategy means that the null hypothesis is tested in a certain order, until the relevant conditions are met and the test is stopped, just like a series relationship, that is, whether or not to perform subsequent hypothesis testing based on the results of the previous hypothesis testing. The order of hypothesis testing and the corresponding multiplicity adjustment methods in sequential strategy have different impacts on the overall conclusions, hence special attention should be paid during the study design stage.

4.1.3 Staged overall decision-making strategy

Staged overall decision-making strategy means that the overall decision-making is carried out in the stages according to the prespecified sequence, a typic example is an interim analysis to show efficacy. An overall decision is made at each stage to determine whether the trial is terminated early due

to superiority or continue due to futility. At each stage, we can use a parallel strategy or sequential strategy in the decision-making strategy to deal with the multiplicity issue. Multistage decision-making requires multiplicity adjustment, that is, each stage will consume a certain amount of α, and the nominal test level α_i at each stage can be either the same or different, depending on the α consumption strategy used.

4.2 Multiplicity adjustment method

The method of multiplicity adjustment is to control the FWER at test level α by adjusting the nominal test level α_i for each independent hypothesis test in the overall decision. The determination of the nominal test level can be chosen according to the decision strategy of the multiplicity problem.

4.2.1 Multiplicity adjustment method based on parallel strategies

(1) Bonferroni method. The basic idea of the Bonferroni method is that the sum of the nominal levels of each individual independent test α_i, is equal to the FWER level α, i.e.

$$\alpha_1+\alpha_2+\ldots+\alpha_i+\ldots+\alpha_m=\alpha$$

Each nominal level can be the same ($\alpha_i=\alpha/m$) or different, and the latter is often used when the importance of each hypothesis testing is different. For instance, if a clinical trial has 3 primary endpoints, 3 hypothesis tests should be performed, with α = 0.05. If the three primary endpoints are equally important, α_i can set to be the same for each test, α_i = 0.0167 (= 0.05/3), and if the *P*-value of one hypothesis test is less than 0.0167, then the corresponding test is considered significant; If the importance of the three primary endpoints is different, then α_1, α_2 and α_3 can be unequal, such as 0.030, 0.015 and 0.005, respectively, then each test is considered to be statistically significant if the *P*-value is less than the corresponding α_i.

(2) Prospective α allocation scheme. The idea of prospective α allocation scheme (PAAS) is close to the Bonferroni method and can be understood

as the product of the complementarities of the nominal test level α_i of each individual hypothesis tests equal to the complementarity of the α, i.e.

$$(1 - \alpha_i)(1 - \alpha_2) \ldots (1 - \alpha_m) \quad = (1- \alpha).$$

Each α_i can be the same or different, if the same, it can be obtained according to the Šidák method

$$\alpha_i=1-(1-\alpha)^{1/m}$$

For example, for a clinical trial with three endpoints, two of which are assigned α_i values, α_1=0.02, α_2=0.025. If α is 0.05, then 0.98 × 0.975 × (1-α_3) = 0.95 according to the above formula, and the α_3 of the third endpoint is 0.0057. If the α_i of the three null hypotheses are assigned equal weights, then α_i based on the Šidák method is found to be 0.01695, it should be noted that the PAAS method can only control FWER when the multiple tests are independent or positively correlated.

4.2.2 Multiplicity Adjustment Methods for Sequential Strategies

(1) Holm method. The Holm method is a modification of the Bonferroni method with progressively smaller test statistics (progressively larger *P*-values). The method first calculates the *P*-values of each hypothesis tested, then ranks the *P*-values in order from small to large as $P_1<P_2<\ldots<P_m$, with its corresponding null hypothesis is H_{01}, H_{02}, …, H_{0m}, and then compare the *P*-values sequentially with the corresponding α_i to test H_{0i} in turn, $1 \leqslant I \leqslant m$. The first step starts with the smallest *P*-value and tests the null hypothesis H_{01}, if $P_1>\alpha_1$ ($=\alpha/m$), then the null hypothesis H_{01} is not rejected and all remaining hypotheses are stopped; if $P_1 \leq \alpha_1$, then H_{01} is rejected and H_{A1} holds and proceed to the next test. The nominal level of the second test is $\alpha_2=\alpha/(m-1)$. Compare the *P*-value of this test with α_2. If $P_2 > \alpha_2$, stop testing the remaining hypothesis; otherwise, H_{A2} is established and proceed to the next test. More generally, when testing the ith null hypothesis $H_{0i,}$ if $P_i>\alpha_i$ ($=\alpha/(m–i+1)$), then stop the test and accept H_{0k}, …, H_{0m}, otherwise,

reject H_{0i} (accept H_{Ai}) and proceed to the next test, and so on

(2) Hochberg method. The Hochberg method is a multiple adjustment method based on the Simes method with progressively larger test statistics (progressively smaller *P*-values). Firstly, the *P*-value of each hypothesis test is calculated, and the *P*-values are sorted from large to small, which is denoted as $P_1>P_2>\ldots>P_m$, and then compared to the corresponding α_i in order of decreasing *P*-value. The first step starts with the largest *P*-value and tests the null hypothesis H_{01}. If $P_1 \leq \alpha_1$ ($=\alpha$), all null hypotheses are rejected and the test is stopped, and accept all alternative hypotheses H_{Ai}; otherwise H_{01} is not rejected and proceed to the next test. The nominal level of the second test is $\alpha_2 = \alpha/2$. The *P*-value of this test is compared with α_2. If $P_2 \leq \alpha/2$, stop testing the remaining hypotheses. Except for H_{A1}, all other alternative hypotheses are true; otherwise, do not reject H_{02} and proceed to the next test. More generally, when testing the ith null hypothesis H_{0i}, if $P_i \leq \alpha_i$ ($=\alpha/i$), stop testing the remaining hypotheses, reject H_{0i}, …, H_{0m}; otherwise, do not reject H_{0i} and proceed to the next test, and so on. It should be noted that the Hochberg method can achieve control of FWER only when all the hypotheses are independent or positively correlated.

(3) Fixed sequence method. Fixed sequence method pre specifies the order in which the hypotheses are tested, the nominal level of all hypothesis test α_i is the same and equals to α. The test begins with the first hypothesis test and the next hypothesis test is performed only if the previous null hypotheses were rejected. The hypothesis test continues until a hypothesis test does not reject the null hypothesis, and the overall conclusion is that all the hypotheses in front of this non-significant hypothesis are statistically significant. For example, there are three null hypotheses in the order H_{01}, H_{02}, and H_{03}. If both the first and second hypotheses reject the null hypothesis at the significant level of α, but the third hypothesis test fails to reject the null hypothesis H_{03}, then both the alternative hypotheses H_{A1} and

H_{A2} are claimed to be true, while H_{A3} is not.

(4) Fallback method. The fallback method is a multiplicity adjustment method with a fixed sequence. It first prespecifies the order in which the hypothesis test is tested and determines the nominal test level α_i for each hypothesis test. Then the hypothesis tests are carried out in that order. H_{01} is tested at α_1 level, and test H_{02} at α_2 level if H_{01} is not rejected. If H_{01} is rejected, H_{02} is tested at the $\alpha_1+\alpha_2$ level, and so on. For example, in a clinical trial with two primary endpoints (O_1 and O_2), the nominal test levels for O_1 and O_2 are assigned to be $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$. Using the fallback method, if the P-value of the hypothetical test for O_1 and O_2 is $P_1 = 0.062$ and $P_2 = 0.005$, respectively, the final conclusion is that the study drug will benefit significantly from O_2 ($P_1 = 0.062 > \alpha_1$, $P_2 = 0.005 < \alpha_2$). If the P-values of the hypothesis test are $P_1 = 0.032$ and $P_2 = 0.015$, respectively, the overall conclusion is that the study drug has significant benefit on both O_1 and O_2 ($P_1 = 0.032 < \alpha_1$, $P_1 = 0.015 < \alpha_1+\alpha_2$).

4.2.3 Common α spending methods for interim analysis

The classical α spending methods in the interim analysis include the Pocock method, O'Brien-Fleming method, and Haybittle-Peto method. A common premise of these three methods is that the proportion of calendar time or cumulative data between the interim analyses is the same. But the assignment of α_i for the hypothesis test in each interim analysis is allowed to be different. A more flexible α spending method is to use the α spending function, such as the Lan-DeMets α spending function, which is an extension of the classical method described above, it is more flexible in selecting interim analysis time points. For example, in a confirmatory clinical trial aimed at evaluating the anti-tumor drugs with immune target inhibitors, the primary endpoint is all-cause death. An interim analysis is planned, and the trial has the possibility to be terminated early based on the efficacy. Because of a possible delay in the onset of action of the immune

target inhibitor, the interim analysis was planned when 75% of the deaths were observed, a relatively late time point in the study. Using the Lan-DeMets α spending function with an approximate the O'Brien Fleming boundary and requiring a two-sided FWER of 0.05, the two-sided nominal test levels were assigned to be 0.019 and 0.044 for the interim and final analyses, respectively.

When the multiplicity issues in a clinical trial are more complicated, a multiplicity adjustment approach with multiple strategies can be jointed used. It is important to note that a simple combination of different multiplicity adjustments does not necessarily control FWER. Therefore, to ensure the control of FWER, the gatekeeping or graphical methods can be considered when multiple multiplicity adjustment approaches are used in combination under complex circumstances.

4.3 Multiplicity analysis method

For the multiplicity issue that needs to be taken care of, most of them are implemented based on the specific statistical analysis methods combined with the multiplicity adjustment method. For example, for multiple endpoints with different data types (such as quantitative, qualitative, and survival time), different statistical analysis methods (such as covariance analysis, Mantel-Haenszel χ^2 test and Kaplan-Meier test) will be adopted for the comparison between groups. Meanwhile, it is also necessary to rely on the multiplicity adjustment method for multiple endpoints (such as the Bonferroni method) to determine the test level α_i for each hypothesis test before the conclusion can be made.

For a single endpoint with multiple group comparisons in the same study, some statistical analysis methods solve the problem of multiple comparisons on the basis of the global hypothesis test. The basic idea is that the standard error involved in pairwise comparison is derived from the standard error of the global hypothesis test. For example, pairwise

comparisons of quantitative outcome variables based on the analysis of variance include LSD method, SNK method, etc., and comparisons of multiple groups with the reference groups include Dunnett method, etc.; for the multiple comparisons of qualitative outcome variables, we can first transform qualitative outcomes to quantitative variables (such as arcsine transformation), and then use the preceding analysis methods for quantitative variables; survival time outcome variables are tested using the log-rank test (Mantel-Cox method) based Kaplan-Meier method, Breslow method (extended Wilcoxon method), etc. It is important to note that some methods do not necessarily control FWER. For statistical analysis methods that cannot achieve multiple comparisons on the basis of global hypothesis tests, local hypothesis tests (pairwise comparisons) combined with α allocation methods (such as the Bonferroni method) should be used to control for FWER.

Using multivariate parametric methods (such as multivariate analysis of variance) is one of the means to solve the problem of multiplicity, especially for the case of multiple endpoints. But such methods generally require that all the endpoints jointly follow the multivariate normal distribution, and the interpretation of the results of such analyses is often not intuitive, which greatly limits their applications.

Repeated sampling (such as the bootstrap method and the permutation method) is also one of the ways to solve the multiplicity problem. The advantage of such methods is that they can control the FWER while preserving a high statistical power; the disadvantage is that the empirical distribution on which it is based is difficult to verify, resulting in insufficient accuracy of the estimate, in addition, the reliability of these methods generally relies on the large sample size. Therefore, such methods are rarely practiced in clinical trials and need to be used with caution. It is recommended to fully communicate with regulatory authorities in advance

before using the resampling-based method for multiplicity adjustment.

Since there are many statistical analysis methods for solving multiplicity issues and each method has its own advantages and disadvantages, the sponsor needs to specify the statistical analysis methods for multiplicity issues in the clinical trial protocol or statistical analysis plan in advance.

5. Other considerations

5.1 Multiplicity issues without adjustment

Situations that do not require multiplicity adjustment include, but are not limited to, the following (none of which includes interim analysis for efficacy):

(1) In clinical trials with multigroup comparisons for a single primary endpoint (for example, the standard three-arm design for a non-inferiority trial), and the new drug is claimed to be effective if all hypothesis tests must be statistically significant;

(2) For the single primary endpoint, the study hypothesis is that the efficacy of the test drug is at least non-inferior to that of the positive control drug, i.e., the hypothesis test is carried out in a fixed order, that is, the hypothesis, H_0, that the efficacy of the study drug was not inferior to that of the positive control drug is tested first. If hypothesis H_0 is rejected, the hypothesis that the efficacy of the study drug was superior to that of the positive control drug is tested in the second step.

(3) For multiple primary endpoints, the study drug is considered effective if and only if the hypothesis testing for all primary endpoints are statistically significant;

(4) For multiple secondary endpoints that do not aim at the benefits claimed in the labeling;

(5) For complex cross-study designs, such as basket design, umbrella design, and platform design, if the sub-study is independent and addresses

the different clinical questions, such as applicable diseases, target population, etc.;

(6) When analyzing data, different analysis data sets may be analyzed for the same primary endpoint, as long as the main analysis data set is defined in advance for the primary efficacy evidence;

(7) Use different statistical models or use different parameter settings for the same model, as long as the main analysis model is defined in advance;

(8) Perform sensitivity analysis according to different assumptions, such as the analysis after imputation with different missing data estimation methods, the analysis with different treatments for outliers, etc.

5.2 Parameter estimation of multiplicity test

The corresponding confidence interval should be estimated according to the multiplicity adjustment method. There are many multiple adjustment methods, some of which are simple and easy to perform interval estimation but relatively conservative. For example, the Bonferroni method can be used to adjust the confidence interval. Some methods are more complicated, and it may be difficult to estimate the corresponding confidence interval.

Multiplicity adjustment also has the potential to introduce a selection bias in point estimates. For example, in confirmatory clinical trials with multiple-dose groups, there is a potential to overestimate the efficacy of a drug if the decision strategy for the multiplicity issue selects the effect size of the dose group that is most different from the placebo in the drug label. Similar selection bias can arise due to the selection of subgroups. Therefore, it is necessary to assess the possible selection bias caused by the multiplicity adjustment.

5.3 Communication with regulatory authorities

The multiplicity issue and the corresponding strategy and method of multiplicity adjustment shall be specified in the clinical trial protocol and statistical analysis plan in advance. For complex multiplicity issues,

whether and how to adjust multiplicity are required to be clearly documented, when the existing strategies and approaches can not address these issues, the sponsors are encouraged to actively communicate with regulatory authorities at the confirmatory clinical trial design stage. During the trial, if major adjustments are made to the clinical trial protocol due to changes in multiplicity adjustment strategies and methods, such changes should be communicated promptly to the regulatory authority.

References

[1] Qian Jun, Chen Pingyan. Multiple comparisons of multiple sample rates. Chinese Journal of Health Statistics, 2008; 25 (2): 206-212.

[2] Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. Statistics in Medicine, 2014; 33 (4): 693-713.

[3] Bretz F, Tamhane AC, Pinheiro J, et al. Multiple Testing in Dose-Response Problem, Chapter 3 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.

[4] Bretz F, Maurer W, Brannath W, et.al. A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 2009; 28 (4): 586-604.

[5] Chen J, Luo JF, Liu K, et al. On power and sample size computing for multiple testing procedures. Computational Statistics and Data Analysis, 2011; 55:(1):110-122.

[6] Collignon O, Gartner C, Haidich AB, et al. Current statistical consultations and regulatory perspectives on the planning of congenital basket umbrella and platform trial. Clinical Pharmacology & Therapeutics, 2020; doi: 10.1002/cpt. 1804.

[7] Dmitrienko A, Tamhane AC, Bretz F, et al. Multiple Testing Methodology, Chapter 2 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.

[8] Dmitrienko A, Tamhane AC, Bretz F, et al. Gatekeeping Procedures in Clinical Trials, Chaper 5 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.

[9] Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 1955; 50 (272): 1096-1121.

[10] European Medicines Agency: Guidance on Multiplicity Islands in Clinical Trials

[11] Freidlin B, Korn EL, Gray R, et.al. Multi-arm clinical trials of new agents: some design considerations. Clinical Cancer Research, 2008; 14 (14): 4368-4371.

[12] Hochberg Y, Tamhane A. Multiplicity comparison procedure. New York: Wiley, 1987.

[13] Howard DR, Brown JM, Todd S, et.al. Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. Statistical Methods in Medical Research, 2018; 27 (5): 1513-1530.

[14] Huque MF, Rohmel J. Multiplicity Problem in Clinical Trials, Chapter 1 of Multiplicity Testing Problem in Pharmaceutical Statistics. CRC Press, 2010.

[15] International Conference on Harmonization (ICH). E9 guideline “Statistical Principles for Clinical Trials”.

[16] International Conference on Harmonization (ICH). E8 guideline “General Considerations for Clinical Trials”.

[17] International Conference on Harmonization (ICH). E17 guideline “General Principles for Planning And Design Of Multi-Regional Clinical Trials”.

[18] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika, 1983; 70 (3):659-663.

[19] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics, 1979; 35 (3):549-556.

[20] Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observations of each patient, I. Introduction and design. British Journal of Cancer, 1976; 34 (6):585-612.

[21] Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika, 1977; 64 (2):191-199.

[22] Sen PK. Some remark on Simes-type multiple tests of significance. Journal of Statistical Planning and Inference, 1999; 82 (1-2):139 – 145.

[23] U.S. Food and Drug Administration. Multiple Endpoints in Clinical Trials – Guidance for the Industry.

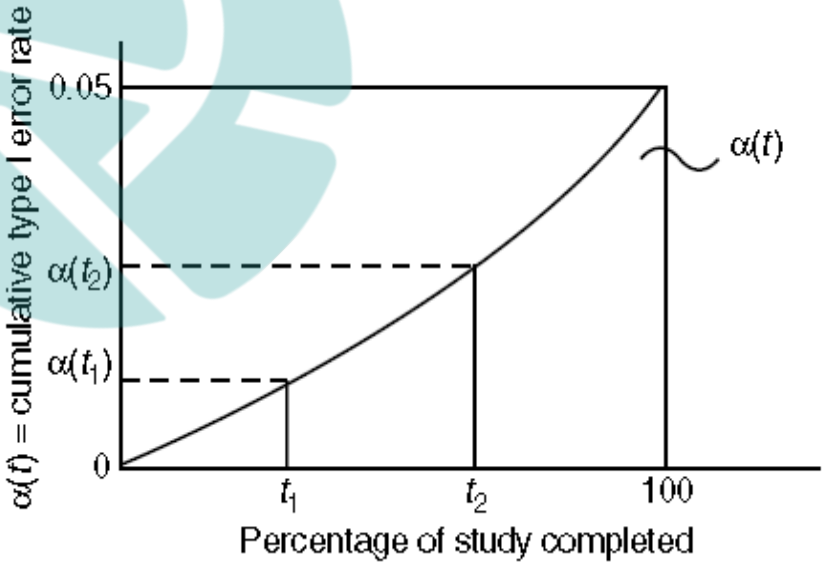
[24] Wang DL, Li YH, Wang X, et al. Overview of multiple testing methodology and recurrent development in clinical trials. Contemporary Clinical Trials, 2015; 45 (Pt A):13-20.

Appendix 1: Glossary

Type I Error: The error in which the null hypothesis is true but the test result rejects it, in clinical trials, it is an error drawing the effectiveness conclusion from statistical inference for an ineffective drug. The probability of making such an error should be controlled at a certain level, which is called test level, or significance level, and is customarily denoted by alpha (α).

Type II Error: The error in which the null hypothesis is incorrect, but the statistical test fails to reject the null hypothesis, in another word, it is the error of statistically drawing an ineffective conclusion on an actually effective drug. This probability is denoted by the symbol beta (β).

***a* Spending Function:** When a clinical study is divided into several stages for overall decision-making (such as interim analysis based on effectiveness), a certain α shall be consumed at each stage. With the progress of the study, the proportion of completed studies (such as *1/3, 1/2, 3/5, etc.)* shows a certain functional relationship with the cumulative type I error rate, as shown in the figure below.



Multiplicity Issues: It refers to the situation in the clinical trial where it

relies on more than one statistical inference (multiple testing) to make decisions on the study conclusions.

Multiplicity Adjustment: A process of controlling the FWER to a reasonable level using appropriate strategies and methods.

Key Secondary Endpoint: The secondary endpoints are used to support the benefits that are claimed in the labeling.

Nominal Level: The test level for a single hypothesis test in the multiple hypothesis tests is known as the nominal test level, also known as the local test level, denoted by α_i.

Familywise Error Rate (FWER): The probability that at least one true null hypothesis is rejected in multiple hypothesis tests of the same test interest, regardless of which null hypothesis or hypotheses are true in multiple tests. It should be controlled at a reasonable level

Primary Endpoint: The endpoint that is directly related to the main problem (primary objective) concerned by clinical trials and can provide the most clinically significant and convincing evidence. It is commonly used for primary analysis, sample size estimation, and evaluation of whether a trial achieves the primary objective.

Appendix 2: Chinese-English Vocabulary

中文	English
α分配	α Allocation
α消耗	α Spending
α消耗函数	α Spending Function
I类错误	Type I Error
II类错误	Type II Error
多重性	Multiplicity
多重性调整	Multiplicity Adjustment
多重性问题	Multiplicity Issue
多个终点	Multiple Endpoints
分题研究	Substudies
关键次要终点	Key Secondary Endpoint
回退法	Fallback Method
剂量-反应关系	Dose-response Relationship
名义检验水准	Nominal Level
前瞻性α分配法	Prospective Alpha Allocation Scheme, PAAS
守门法	Gatekeeping
图示法	Graphical Approach
显著性水准	Significance Level
总I类错误率	Familywise Error Rate, FWER