Multiple Endpoints in Clinical Trials

Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

October 2022
Biostatistics

Multiple Endpoints in Clinical Trials

Guidance for Industry

Additional copies are available from:

Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10001 New Hampshire Ave., Hillandale Bldg., 4th Floor
Silver Spring, MD 20993-0002
Phone: 855-543-3784 or 301-796-3400; Fax: 301-431-6353
Email: druginfo@fda.hhs.gov
https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs

and/or

Office of Communication, Outreach and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave., Bldg. 71, Room 3128
Silver Spring, MD 20993-0002
Phone: 800-835-4709 or 240-402-8010
Email: ocod@fda.hhs.gov
https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

October 2022
Biostatistics

Contains Nonbinding Recommendations

TABLE OF CONTENTS

Contains Nonbinding Recommendations

Multiple Endpoints in Clinical Trials Guidance for Industry[1]

This guidance represents the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA office responsible for this guidance as listed on the title page.

I. INTRODUCTION

This guidance provides sponsors and review staff with the Agency's thinking about the problems posed by multiple endpoints in the analysis and interpretation of study results and how these problems can be managed in clinical trials for human drugs, including drugs subject to licensing as biological products. Most clinical trials performed in drug development contain multiple endpoints to assess the effects of the drug and to document the ability of the drug to favorably affect one or more disease characteristics. When more than one endpoint is analyzed in a single trial, the likelihood of making false conclusions about a drug's effects with respect to one or more of those endpoints could increase if there is no appropriate adjustment for multiplicity. The purpose of this guidance is to describe various strategies for grouping and ordering endpoints for analysis of a drug's effects and applying some well-recognized statistical methods for managing multiplicity within a study to control the chance of making erroneous conclusions about a drug's effects. Basing a conclusion on an analysis where the risk of false conclusions has not been appropriately controlled can lead to false or misleading representations regarding a drug's effects.

The ICH guidance for industry *E9 Statistical Principles for Clinical Trials* (September 1998)[2] is a broad ranging guidance that includes discussion of multiple endpoints. This guidance on multiple endpoints in clinical trials for human drugs provides greater detail on the topic. The issuance of this guidance represents partial fulfillment of an FDA commitment under the Food and Drug Administration Amendments Act (FDAAA) of 2007.

[1] This guidance has been prepared by the Office of Biostatistics in the Office of Translational Sciences in the Center for Drug Evaluation and Research in cooperation with the Center for Biologics Evaluation and Research at the Food and Drug Administration.

[2] The ICH E9 guidance is available on the FDA guidance web page under the topic ICH – Efficacy. We update guidances periodically. For the most recent version of a guidance, check the FDA guidance web page at https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. BACKGROUND AND SCOPE

Efficacy endpoints are measures designed to reflect the intended effects of a drug. They include assessments of clinical events (e.g., mortality, stroke, pulmonary exacerbation, venous thromboembolism), symptoms (e.g., pain, dyspnea, symptoms of depression), measures of function (e.g., ability to walk or exercise), or surrogate endpoints that are reasonably likely or expected to predict a clinical benefit.

Because most diseases can potentially cause more than one clinical event, symptom, and/or altered function, many trials are designed to examine the effect of a drug on more than one aspect of the disease. In some cases, efficacy cannot be adequately established based on a single disease aspect, and the study should use either an endpoint that incorporates multiple aspects of the disease into a single endpoint or effects should be demonstrated on multiple endpoints. In other cases, an effect on any of several endpoints could be sufficient to support approval of a marketing application.

Failure to account for multiplicity when there are several endpoints evaluated in a study can increase the chance of false conclusions regarding the effects of the drug. The regulatory concern regarding multiplicity arises principally in the evaluation of clinical trials intended to demonstrate effectiveness supporting drug approval and claims in FDA-approved labeling; however, this issue is important for trials throughout the drug development process. For instance, if safety outcomes are to be assessed via hypothesis testing, they would be subject to the multiplicity considerations described in this guidance. Multiplicity problems for safety analyses that are not part of a prespecified set of hypotheses for formal statistical testing are outside the scope of this guidance.

In the following sections, the issues of multiple endpoints and methods to address them are discussed. The issues of multiplicity and methods that apply to multiple endpoints also generally apply to other sources of multiplicity, including other estimand[3] attributes (e.g., multiple doses, time points, or study population subgroups); however, these other sources of multiplicity will not be specifically addressed in this guidance. Furthermore, there may be different considerations related to multiplicity in certain unique settings, such as the evaluation of multiple different drugs for a single disease in a master protocol, that are not addressed in this guidance. This guidance focuses on the analysis and interpretation of multiple endpoints within a single clinical trial.

[3] See the ICH Guidance for Industry *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials* (May 2021).

A. Demonstrating the Study Objective of Effectiveness

A conclusion that a study has demonstrated an intended effect of a drug is critical to meeting the legal standard for substantial evidence of effectiveness required to support approval of a new drug (i.e., "… adequate and well-controlled investigations…on the basis of which it could fairly and responsibly be concluded…that the drug will have the effect it purports…to have…") (section 505(d) of the FD&C Act).[4] FDA regulations further establish that to be adequate and well controlled, a clinical study of a drug must include, among other things, "an analysis of the results of the study adequate to assess the effects of the drug," a requirement that furthers the "purpose of conducting clinical investigations of a drug," which is "to distinguish the effect of a drug from other influences, such as spontaneous change in the course of the disease, placebo effect, or biased observation."[5] There are also other important factors (e.g., clinical relevance of the endpoint and estimated effect, relevant external information) that are considered in evaluating substantial evidence of effectiveness beyond the results of hypothesis tests in a single trial. A more general discussion of demonstrating substantial evidence of effectiveness can be found in other FDA guidance documents[6] and is outside the scope of this document.

Hypothesis testing is commonly used to address the uncertainty in the assessment of a treatment effect on a chosen endpoint. This approach begins with stating the relevant hypotheses for a chosen endpoint. In the simplest situation where the aim is to demonstrate the superiority of a test drug over control, two mutually exclusive hypotheses are specified for the endpoint in advance of conducting a clinical trial:

- One hypothesis, the null hypothesis, states that there is no treatment effect on the chosen endpoint.

- The other hypothesis is called the alternative hypothesis and posits that there is at least some treatment effect of the test drug.

This pair of hypotheses are tested using a prespecified statistical test to determine whether the trial results are sufficiently unlikely under the null hypothesis so that the null hypothesis can be rejected in favor of the alternative hypothesis. Note that if the null hypothesis is not rejected, it does not necessarily mean that the null hypothesis is true. There are many other potential reasons that could lead to a failure to reject the null hypothesis, such as insufficient sample size.

---

[4] See 21 U.S.C. 355. Biological products are licensed based on a demonstration of safety, purity, and potency (section 351(a)(2)(C) of the Public Health Service Act, 42 USC 262(a)(2)(C)). Potency has long been interpreted to include effectiveness (21 CFR 600.3(s)). In 1972, FDA initiated a review of the safety and effectiveness of all previously licensed biological products. The Agency stated then that proof of effectiveness would consist of controlled clinical investigations as defined in the provision for adequate and well-controlled studies for new drugs (21 CFR 314.126), unless waived as not applicable to the biological product or essential to the validity of the study when an alternative method is adequate to substantiate effectiveness." (37 FR 16681, August 18, 1972).

[5] See 21 CFR 314.126(b)(7), 314.126(a).

[6] See the FDA draft guidance for industry *Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products* (December 2019). When final, this guidance will represent the FDA's current thinking on this topic.

Sometimes (e.g., in some vaccine trials), demonstration of an effect of at least some minimum size is considered critical for approval of a drug. In this case, if formal statistical testing is used for the demonstration, the null hypothesis might be modified to incorporate the smallest clinically meaningful effect that could be accepted.

This guidance focuses on a statistical framework based on hypothesis testing. Sponsors should discuss early with FDA plans to use other approaches (e.g., Bayesian approaches) for a specific development program such as for pediatrics.

B. Type I Error

The rejection of the null hypothesis supports the study conclusion that there is a difference between treatment groups but does not constitute absolute proof that the null hypothesis is false. There is always some possibility of mistakenly rejecting the null hypothesis when it is, in fact, true. Such an erroneous conclusion is called a Type I error. For an endpoint, the probability of falsely rejecting its null hypothesis and, thus, concluding that there is a treatment effect due to the drug on this endpoint when, in fact, there is none, is called the Type I error probability or Type I error rate for this endpoint. The significance level, denoted as alpha (α), is the threshold below which the Type I error rate should be controlled. Null hypothesis rejection is based on a determination that the probability of observing a result at least as extreme as the result of the study assuming the null hypothesis is true (the p-value) is sufficiently low (usually no larger than α).

The alternative hypothesis can be one-sided or two-sided, and statistical tests are performed accordingly. For two-sided hypothesis statistical tests, the Type I error probability refers to the probability of concluding that there is a difference (beneficial or harmful) between the drug and control when there is no difference. For one-sided hypothesis tests, the Type I error probability refers to the probability of concluding specifically that there is a beneficial difference due to the drug when there is not. The most widely used values for α are 0.05 for two-sided tests and 0.025 for one-sided tests. In the case of two-sided tests, an α of 0.05 means that the probability of falsely concluding that the drug differs from the control in either direction (benefit or harm) when no difference exists is no more than 5%, or 1 chance in 20. In the case of one-sided tests, an α of 0.025 means that the probability of falsely concluding a beneficial effect of the drug when none exists is no more than 2.5%, or 1 chance in 40. Use of a two-sided test with an α of 0.05 that allocates the α symmetrically to each side generally also ensures that the probability of falsely concluding benefit when there is none is no more than approximately 2.5% (1 chance in 40). These Type I error rates are correct if the statistical test is appropriate. If there are issues with the statistical test (e.g., the underlying assumptions do not hold), the Type I error rate could be even larger.

FDA's concern for controlling the Type I error probability is to minimize the chances of a false favorable conclusion for any primary or secondary endpoints (see section III.), regardless of which and how many of these endpoints in the study have no effect. The Type I error probability associated with testing multiple endpoints of a study is called overall Type I error probability. The rationale for controlling this probability is given in the next subsection (section II.C.). When

there is more than one primary or secondary endpoint, it is important to ensure that the evaluation of multiple hypotheses will not lead to inflation of the study's overall Type I error probability (or rate) relative to the planned level. To control the Type I error rate, it is critical that sponsors prospectively specify the following:

- all endpoints in the primary and secondary families (see section III. for definitions).
- all data analyses that will be performed to test hypotheses about the prespecified endpoints, regardless of whether they are considered primary or secondary.

For a study with multiple endpoints, the analysis plan should describe the testing procedure for the hypotheses being tested with a proper control of overall Type I error rate.

C. Multiplicity

In a clinical trial with a single endpoint tested at two-sided $\alpha = 0.05$, the probability of finding a difference between the treatment group and a control group in favor of the treatment group when no difference exists in the population is 0.025 (a 2.5% chance). That is, there is a 97.5% chance of appropriately not finding a favorable effect if there is no true effect for this endpoint. By contrast, if there are two independent endpoints, each tested at two-sided $\alpha = 0.05$, and if success on either endpoint by itself would lead to a conclusion of a drug effect, the chance of appropriately not finding a favorable effect on both endpoints together is thus 0.975 * 0.975, which is approximately 0.95, and so the probability of falsely finding a favorable effect on at least one endpoint is approximately 0.05. Thus, the overall Type I error rate in favor of the drug nearly doubles when two independent endpoints are tested. This higher-than-intended overall Type I error rate when multiple tests are conducted without adjustment is called the multiplicity problem. Thus, without correction for multiplicity, the chance of making a Type I error for this example study as a whole would rise to approximately as high as 5% in favor of the drug, and, therefore, the overall Type I error rate would not be adequately controlled. The problem is exacerbated when more than two endpoints are considered. For example, for three independent endpoints, the Type I error rate is 1 - (0.975 * 0.975 * 0.975), which is about 7%. For ten independent endpoints, the Type I error rate is about 22%. If the multiple endpoints are correlated, the overall Type I error rate is also inflated but potentially by a lesser degree.

Even when a single outcome variable is being assessed, if multiple facets of that outcome are analyzed (e.g., multiple dose groups, multiple time points, or multiple subject subgroups based on demographic or other characteristics) and if any one of the analyses is used to conclude that the drug has been shown to produce a beneficial effect, the multiplicity of analyses may cause inflation of the Type I error rate. Hence, by inflating the Type I error rate, multiplicity produces uncertainty in interpretation of the study results such that the conclusions about whether effectiveness has been demonstrated in the study become unreliable. There are various approaches that can be planned prospectively and applied to maintain the overall Type I error rate at 2.5% or below.

For controlling multiplicity, an important principle is to first prospectively specify all planned endpoints, time points, analysis populations, doses, and analyses; then, once these factors are

specified, appropriate adjustments for multiple endpoints and analyses can be selected, prespecified, and applied, as appropriate. Changes in the analytic plan to perform additional analyses can reintroduce a multiplicity problem that can negatively impact the ability to interpret the study's results unless these changes are made prior to data analysis and appropriate multiplicity adjustments are performed. The statistical analysis plan should not be changed after unmasking of treatment assignments and performing statistical analyses.

A focus of this guidance is control of the Type I error rate for the prespecified set of endpoints (i.e., primary and secondary endpoints) of a clinical trial to ensure that the major findings of a clinical trial are well supported, and the effects of the drug have been demonstrated. Analyses that explicate the characteristics of an effect on an endpoint that has been demonstrated—such as time of onset, distribution of effect sizes across the population, effects in subgroups, and effects on the components of a composite endpoint—are all descriptive to provide a deeper understanding of the nature of that endpoint finding, and do not extend to effects outside of that endpoint. These descriptive analyses can be considered for inclusion in the FDA-approved labeling without presenting p-values.

Of note, there is not always a clear-cut distinction between an analysis closely related to a major finding and one that demonstrates additional effects. Therefore, when definitive conclusions are to be drawn, such analyses should be prespecified and appropriately included in the prespecified multiple-testing strategy. A descriptive analysis that is not included in the prespecified multiple-testing strategy should not be presented in FDA-approved labeling in ways that imply a statistically rigorous conclusion or convey certainty about the effects that are not supported by that trial. Descriptive analyses are not the subject of this guidance and are not addressed in detail.

III. MULTIPLE ENDPOINTS: GENERAL PRINCIPLES

A. The Hierarchy of Families of Endpoints

Endpoints in adequate and well-controlled drug trials are usually grouped hierarchically, often according to their clinical importance, but also taking into consideration the expected frequency of the endpoint events and anticipated drug effects. The critical determination for grouping endpoints is whether they are intended to establish effectiveness to support approval or intended to demonstrate additional meaningful effects. Endpoints critical to establish effectiveness for approval are often designated as primary endpoints. Secondary endpoints can provide useful description to support the primary endpoint(s) and/or demonstrate additional clinically important effects. The third category in the hierarchy includes all other endpoints, which are referred to as exploratory. Exploratory endpoints can include endpoints for research purposes or for new hypotheses generation. Each category in the hierarchy can contain a single endpoint or a family of endpoints.

1. Primary Endpoint Family

The endpoint(s) that establish the effect(s) of the drug and will be the basis for concluding that the study meets its objective are designated the primary endpoint family. When there is a single

prespecified primary endpoint, there are no multiple-endpoint-related multiplicity issues in the determination that the study achieves its objective.

Multiple primary endpoints occur in three ways, further described in section III.C. The first is when there are multiple primary endpoints, and each endpoint could be sufficient on its own to establish the drug's efficacy. These multiple endpoints thus correspond to multiple chances of success, and in this case, failure to adjust for multiplicity can lead to Type I error rate inflation and a false conclusion that the drug is effective. The second is when the determination of effectiveness depends on success on all primary endpoints, when there are two or more primary endpoints. In this setting, there are no multiplicity issues related to primary endpoints, as there is only one path that leads to a successful outcome for the trial and therefore, no concern with Type I error rate inflation. In the third, critical aspects of effectiveness can be combined into a single primary composite or other multicomponent endpoint, thereby avoiding multiple-endpoint-related multiplicity issues. For example, in many cardiovascular studies it is usual to combine several endpoints (e.g., cardiovascular death, heart attack, and stroke) into a single composite endpoint that is primary and to consider death a secondary endpoint (see section III.A.2.).

2. Secondary and Exploratory Endpoint Families

When an effect on the primary endpoint is shown, the secondary endpoints can be formally tested. A secondary endpoint could be a clinical effect related to the primary endpoint that extends the understanding of that effect (e.g., an effect on survival when a cardiovascular drug has shown an effect on the primary endpoint of heart failure-related hospitalizations) or provide evidence of a clinical benefit distinct from the effect shown by the primary endpoint (e.g., a disability endpoint in a multiple sclerosis treatment trial in which relapse rate is the primary endpoint). As a general principle, it is important to include the secondary endpoints that can potentially provide evidence of additional effects of the drug on the disease or condition in the Type I error control plan.

In general, it may be desirable to limit the number of secondary endpoints, because if multiplicity adjustments are used, the chance of demonstrating an effect on any secondary endpoint may become increasingly small as the number of secondary endpoints increases, or if a hierarchy is used, the important hypotheses further down the hierarchy might never get tested.

Exploratory endpoints do not need multiplicity adjustment because they are generally not used to support conclusions.

3. Selecting and Interpreting the Endpoints in the Primary and Secondary Endpoint Families

Positive results on the secondary endpoints can be interpretable if there is first a demonstration of a treatment effect on the primary endpoint family (O'Neill 1997). The overall Type I error rate should control for the primary and secondary endpoint families all together.

Occasionally, there are trials where a clinically important endpoint (e.g., mortality or irreversible morbidity) is expected to have too few events to provide adequate power for the trial, while a

different clinically important endpoint occurs more frequently or earlier in the disease process, leading to larger power. In such cases, generally the endpoint with inadequate power for detection is classified as a secondary endpoint, while the endpoint for which larger power is expected is classified as the primary endpoint. For example, in some oncology trials, progression-free survival is selected as the primary endpoint, and overall survival is selected as the secondary endpoint because an effect of treatment on disease progression is clinically important and may be more readily demonstrable, may be detected earlier, and may often be larger because the observed effect on overall survival can be impacted by subsequent treatment post progression.

B. Type II Error Rate and Sample Size

FDA is also concerned with the risk of making a Type II error, which is failing to show an effect of a drug where there actually is one. The study power is the probability that the study will be successful if a treatment effect of a specified size is in fact present. The desired power is an important factor in determining the sample size, especially for the primary endpoints.

The sample size of a study is generally chosen to provide a reasonably high power to show a treatment effect if an effect of a specified size on the primary endpoint(s) is in fact present. The sample size calculation may need to account for the statistical adjustments to control the Type I error rate for multiplicity. For example, if a lower α level is used for a study endpoint, then the sample size should be adjusted to provide desired statistical power for this endpoint.

Using two or more endpoints for which demonstration of an effect on each is recommended to support regulatory approval (called co-primary endpoints; see section III.C.1. below) will increase the Type II error rate and decrease study power. For example, assume two endpoints have the same effect size and the study sample size is selected to provide 80% power to show success on each of these two endpoints. If the endpoints are independent, the power to show success on both will be approximately 64% (0.8 x 0.8); i.e., the likelihood of the study failing to support a conclusion of a favorable drug effect when such an effect existed (the Type II error rate) would be 36%. To maintain desired study power, a larger sample size is recommended, and the individual endpoints could be powered at approximately 90% to ensure the probability of success is at least 80%. The calculation would be different if the endpoints were highly positively correlated or the power was not equal for each endpoint.

C. Types of Multiple Endpoints

Multiple endpoints can be used when demonstration of a drug effect on more than one disease aspect or outcome is critical for determining that the drug confers a clinical benefit. Multiple endpoints can also be used when (1) there are several important aspects of a disease or several ways to assess an important aspect, (2) it may not be known in advance which aspect is more likely to show a drug effect, and (3) an effect on any one endpoint will be sufficient as evidence of effectiveness to support approval. In some cases, multiple aspects of a disease can appropriately be combined into a single endpoint, but subsequent analysis examining each disease aspect or component of this endpoint is generally important for an adequate understanding of the drug's effect. These circumstances are discussed in more detail below.

1. When Demonstration of Treatment Effects on Two or More Distinct Endpoints Is Recommended to Establish Clinical Benefit (Co-Primary Endpoints)

For some disorders, there are two or more different features that are so critically important to the disease under study that a drug will not be considered effective without demonstration of a treatment effect on all of these disease features. The term used in this guidance to describe this circumstance of multiple primary endpoints is co-primary endpoints. Multiple primary endpoints become co-primary endpoints when demonstrating an effect on each of the endpoints is critical to concluding that a drug is effective.

Therapies for the acute treatment of migraine headaches illustrate this circumstance. Although pain is the most prominent feature, migraine headaches are also characterized by the presence of photophobia, phonophobia, and/or nausea, all of which are clinically important. Which of the three is most clinically important varies among individuals. An approach to studying acute treatments for migraine headaches is to consider a drug effective for migraines only if the proportion of subjects with no headache pain at 2 hours after dosing and the proportion of subjects with absence of the most bothersome associated symptom at 2 hours after dosing are both shown to be improved by the drug treatment. Another approach could be to evaluate the drug effect on a response endpoint where response is defined by the absence of both pain and an individually specified second symptom within an individual subject. This approach would utilize a single multi-component endpoint rather than co-primary endpoints.

Trials of combination vaccines are a situation in which co-primary endpoints are applicable. These vaccine trials are typically designed and powered for demonstration of a successful outcome on effectiveness endpoints for each pathogen against which the vaccine is intended to provide protection.

As discussed in section III.B., there is no multiplicity problem when the study is designed to demonstrate efficacy on all of the separate endpoints. However, co-primary endpoint testing increases the Type II error rate. In general, unless clinically very important, the use of more than two co-primary endpoints should be carefully considered because of the loss of power.

There have been suggestions that the statistical testing criteria for each co-primary endpoint could be increased (e.g., testing at an α of 0.06 or 0.07) when the targeted α is 0.05 to accommodate the loss in statistical power arising from the need to show an effect on both endpoints. Increasing α for each co-primary endpoint is not acceptable because doing so may undermine the ability to interpret a treatment effect on each disease aspect considered critical to show that the drug is effective in support of approval.

2. When Demonstration of a Treatment Effect on at Least One of Several Primary Endpoints Is Sufficient

Many diseases have multiple sequelae, and an effect demonstrated on any one of these aspects could support a conclusion of effectiveness. Selection of a single primary endpoint may be difficult, however, if the aspect of a disease that will be responsive to the drug or the evaluation

method that will better detect a treatment effect is not known a priori (at the time of trial design). In this circumstance, a study might be designed such that success on any one of several endpoints could support a conclusion of effectiveness. This creates a primary endpoint family. For example, consider a drug for the treatment of burn wounds where it is not known whether the drug will increase the rate of wound closure or reduce scarring, but the demonstration of either effect alone would be considered clinically important. A study in this case might have both wound closure rate and a scarring measure as separate primary endpoints.

This use of multiple endpoints creates a multiplicity problem because there are several ways for the study to successfully demonstrate a treatment effect. Control of the Type I error rate for the primary endpoint family is critical. A variety of approaches can be used to address this multiplicity problem; the appendix describes and discusses some of these approaches.

3. Composite Endpoints

There are some disorders for which more than one clinical outcome in a clinical trial is important, and all outcomes are expected to be affected by the treatment. Rather than using each as a separate primary endpoint (creating multiplicity) or selecting just one to be the primary endpoint and designating the others as secondary endpoints, it could be appropriate to combine those clinical outcomes into a single variable. This is often called a composite endpoint, where an endpoint is defined as the occurrence or realization in a subject of any one of the specified components. A typical example is a composite of major adverse clinical outcome events in cardiovascular trials (e.g., a composite of myocardial infarction, stroke, or death). When the components correspond to distinct events, composite endpoints are often assessed as the time to first occurrence of any one of the components. If a single statistical test is performed on the composite endpoint, no multiplicity problem will occur for this endpoint.

One possible reason for using a composite endpoint is that the incidence of each of the events may be too low to allow a study of reasonable size to have adequate power; the composite endpoint can provide a substantially higher overall event rate that allows a study with a reasonable sample size and study duration to have adequate power. Composite endpoints are often used when the goal of treatment is to prevent or delay occurrence of one of several clinically important and related events (e.g., use of an anti-platelet drug in subjects with coronary artery disease to prevent myocardial infarction, stroke, or death), possibly without knowledge of which event(s) may be affected.

The choice of the components of a composite endpoint should be made carefully. The treatment effect on the composite event rate can be interpreted as characterizing the overall clinical effect when the individual events all have reasonably similar clinical importance. The effect on the composite endpoint, however, will not be a reasonable indicator of the effect on all of the components or an accurate description of the drug's benefit if the clinical importance of different components is substantially different and the treatment effect is chiefly on the least important event. Furthermore, it is possible that a component with greater importance would be adversely affected by the treatment, even if one or more event types of lesser importance are favorably affected, so that although the overall outcome still has a favorable statistical result, doubt may arise about the treatment's clinical value. In this case, although the overall statistical analysis

indicates the treatment is beneficial, careful examination of the data could call this conclusion into question. For this reason, as well as for a greater depth of understanding of the treatment's effects, analyses of the components of the composite endpoint are important (see section III.D.) and can influence interpretation of the overall study results. The examination of the components is always necessary, but whether multiplicity adjustment should be made depends on the purpose. If the intent is to better understand the demonstrated effect on the composite, then no adjustment is recommended. In that case, clinical judgment is used to decide whether the benefit is clinically meaningful and exceeds risk, and how it will be described in the FDA-approved labeling. If the intent is to establish additional effects of the drug, then multiplicity adjustment should be made.

4. Multi-Component Endpoints

A multi-component endpoint is a within-subject combination of two or more components. In this endpoint, an individual subject's evaluation is dependent upon observation of all the specified components in that subject. A single overall rating or status is then determined according to specified rules.

A single overall rating can be formed by some kind of average (either weighted or unweighted) across the individual domain scores. An example of a multi-component endpoint is the Positive and Negative Syndrome Scale (PANSS) in schizophrenia research. A multi-component endpoint can also be a dichotomous (response) endpoint corresponding to an individual subject achieving specified criteria on each of the multiple components. For example, the primary endpoint in clinical trials of allogeneic pancreatic islet cells for Type 1 diabetes mellitus can be a response rate in which subjects are considered responders only if they meet two dichotomous response criteria: normal range of HbA1c and elimination of hypoglycemia.

There are more complex endpoint formulations where several, but not all, different features of a disease must be positively affected for a subject to be regarded as receiving benefit. For example, a positive response for an individual subject might be defined as a certain degree of improvement in two specific aspects of a disease along with improvement in at least three out of five additional disease features, as in the American College of Rheumatology (ACR) scoring system for rheumatoid arthritis.

The use of within-subject multi-component endpoints may be efficient if the treatment effects on the different components are generally trending in the same direction within a subject. Study power can be adversely affected, however, if there is limited concordance among the endpoints. Although multi-component endpoints can provide some gains in efficiency compared to co-primary endpoints, the appropriateness of a particular within-subject multi-component endpoint is generally determined by clinical, rather than statistical, considerations. Similar to the assessment of the component endpoints of a composite endpoint in section III.C.3., evaluation of the components of a multi-component endpoint may be important but should be subject to pre-specification and multiplicity adjustment if the intent is to support specific conclusions on how a treatment affects specific components (see section III.D.).

5. Clinically Critical Endpoints Too Infrequent for Use as a Primary Endpoint

For many serious diseases, there is an endpoint of such great clinical importance that it is unreasonable not to collect and analyze the endpoint data; the usual example is mortality or major morbidity events (e.g., stroke, fracture, pulmonary exacerbation). Even if relatively few of these events are expected to occur in the trial, they can be included in a composite endpoint (see section III.C.3.) and also designated as a planned secondary endpoint to potentially support a conclusion regarding effect on that separate endpoint, if the effect of the drug on the composite primary endpoint is demonstrated.

D. The Individual Components of Composite and Multi-Component Endpoints

1. Evaluating and Reporting the Results of Composite Endpoints

For composite endpoints whose components correspond to events, an event is usually defined as the first occurrence of any of the designated component events. Such composites can be analyzed either with comparisons of proportions between study groups at the end of the study or using time-to-event analyses. The time-to-event method of analysis is the more common method when, within the study's timeframe of observation, the duration of being event-free is clinically meaningful. Although there may be an expectation that the drug will have a favorable effect on all the components of a composite endpoint, that is not a certainty. Results for each component event should therefore be individually examined and should be included in study reports. These analyses will not alter a conclusion about the statistical significance of the composite primary endpoint; however, interpretation of the result of the composite endpoint can be uncertain (see section III.C.3.). If there is an interest in analyzing one or more of the components of a composite endpoint as distinct hypotheses to demonstrate effects of the drug, the hypotheses should be part of the prospectively specified statistical analysis plan that accounts for the multiplicity this analysis will entail, as described above, for mortality. However, testing for individual component endpoints is likely to be underpowered as the sample size or total number of events is usually planned for testing the composite endpoint.

Decomposition of the first composite event is often presented to depict how the component events constitute the composite event in terms of proportion. For example, in the RENAAL trial (Brenner et al. 2001), the primary efficacy endpoint was the first occurrence of the composite endpoint of doubling of serum creatinine, end-stage renal disease, or death. Based on such decomposition, 52% of the first composite events were doublings of serum creatinine, 19% were end-stage renal disease events, and 29% were deaths. However, subjects may experience more than one event type. For these subjects, events occurring after the first composite event (e.g., end-stage renal disease or death occurring after a doubling of serum creatinine) would not be counted in the decomposition. Therefore, evaluation of the individual event types in analyses that include all events for the event type of interest (even those that occur after events of other event types) is also important. Such analyses could demonstrate a possible additional effect of the drug if they are pre-specified, multiplicity is properly accounted for, and the results are interpretable.

2. Evaluating and Reporting the Results on Other Multi-Component Endpoints

As with composite endpoints, understanding which components of a within-subject multi-component endpoint have contributed most to the overall statistical significance could be important to correctly understanding the clinical effects of the drug. Consequently, analysis of the study results on the individual components is usually important but, as stated previously, if undertaken, should not be presented in FDA-approved labeling in ways that imply a statistically rigorous conclusion or convey certainty about the effects that are not supported by that trial. For many of these multi-component endpoints, the overall score is regarded as comprehensive and clinically interpretable. The individual component scales, however, may or may not be independently clinically interpretable. Analyses of specific components or subdomains of a clinical outcome assessment as explicit endpoints in the primary or secondary endpoint families can be reasonable, contingent on the endpoint being clinically interpretable. Pre-specification of specific components or subdomains as endpoints with appropriate multiplicity control is recommended if the intent is to demonstrate an effect of a drug on one or more of these endpoints in addition to the overall multi-component endpoint.

IV. METHODOLOGICAL CONSIDERATIONS

A variety of situations in which multiplicity arises have been discussed in sections II. and III. When there is a family of endpoints (discussed in section III.A.), the probability of erroneously finding a statistically significant treatment effect in at least one endpoint regardless of the presence or absence of treatment effects in the other endpoints is the overall Type I error rate. This error rate is typically held to 0.05 (or 0.025 for one-sided tests). Statistical methods that control this error rate at the desired level can permit an effectiveness conclusion on individual endpoints.

There are many common statistical methods for addressing multiple-endpoint-related multiplicity problems (Hochberg and Tamhane 1987). The appendix presents some of the commonly considered methods. Examples include the Bonferroni, Holm (Holm 1979), and Hochberg (Hochberg 1988) procedures, which do not assume any hierarchy among the tested null hypotheses (i.e., any individual null hypothesis in the family can be rejected regardless of the rejection of other hypotheses). Other viable methods apply a combination of partial alpha allocation and hierarchies, such as graphical methods (Bretz et al. 2009) that are presented in the appendix. If finding a statistically significant treatment effect in any one of the considered endpoints is considered a success, then methods that appropriately adjust for multiplicity across the family of endpoints can be applicable.

However, if endpoints are ordered based on clinical importance or logically related, then different methods can be recommended (e.g., Pocock et al. 2012). For example, in the simple case where there is one primary and one secondary endpoint, a hierarchical testing approach can be used. Some methodologies have been developed to account for more complex logical/ hierarchical relationships among the endpoints such as graphical approaches (e.g., Bretz et al. 2009) and mixture gatekeeping procedures (Dmitrienko et al. 2008). The graphical method has a sequential testing algorithm and makes it possible to visualize the testing process via a graph.

In some cases, a primary endpoint can be tested for non-inferiority (with a fixed margin), followed by testing it for superiority. If this endpoint is the only endpoint being tested, then non-inferiority and superiority can be tested without multiplicity adjustment because the null hypotheses of non-inferiority and superiority are naturally ordered, and the two tests apply to the one hierarchy considered for this endpoint. However, if at least one more endpoint is included for testing, then multiplicity issues arise, and adjustments should be made to control the overall Type I error probability. For example, the tests could be ordered in a single hierarchy where the additional endpoint(s) are tested after the superiority hypothesis for the primary endpoint. Or, alternatively, testing could proceed to both the superiority hypothesis for the primary endpoint and to the hypotheses for the additional endpoints, with alpha allocation across these multiple hypotheses. To see why such alpha allocation can be applicable, suppose the drug is non-inferior to the active control with respect to the primary endpoint, but the drug is neither superior to the active control for the primary endpoint nor non-inferior to the active control for the secondary endpoint. Thus, a Type I error could occur with either of these hypothesis tests. If both of these were tested at 0.05, the probability of at least one of these leading to a spurious conclusion would be greater than 0.05. Thus, there should be appropriate control in some manner (e.g., test the secondary endpoint only if the primary endpoint superiority is shown or split alpha between the two tests). Additional discussion on this special case and on other methodological considerations is provided in the appendix.

V. SUMMARY

Making a false positive conclusion about effectiveness (i.e., falsely concluding that a drug has a positive treatment effect when it does not) is a major concern. A common approach is to control the Type I error rate at less than 5% (1 in 20 chance) for a false conclusion that there is a treatment difference or 2.5% (1 in 40 chance) for a false positive conclusion about effectiveness. As the number of endpoints or analyses increases, the Type I error rate can increase well beyond 2.5% due to multiplicity. Multiplicity adjustments, as described in this guidance, provide means for controlling the Type I error rate when the drug effect is evaluated in multiple endpoints. There are many strategies and methods that can be used, as appropriate, as described in this guidance. Each of these methods has advantages and disadvantages, and the selection of suitable strategies and methods is a challenge that should be addressed at the study-planning stage. Statistical expertise should be enlisted to help choose the most appropriate approach. Failing to appropriately control the Type I error rate may increase the risk of a false positive conclusion; this guidance is intended to clarify when and how multiplicity due to multiple endpoints should be managed to avoid reaching such false conclusions.

VI. GENERAL REFERENCES

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

Bauer P, 1991, Multiple testing in clinical trials, Statistics in Medicine, 10: 871-890.
Bretz, F, T Hothorn, and P Westfall, Multiple Comparisons Using R, Boca Raton (FL): Chapman and Hall/CRC.

Brenner BM, ME Cooper, D de Zeeuw, WF Keane, WE Mitch, H-H Parving, G Remuzzi, SM Snapinn, Z Zhang, and S Shahinfar, 2001, Effects of Losartan on Renal and Cardiovascular Outcomes in Patients with Type 2 Diabetes and Nephropathy, New England Journal of Medicine, 345:861-869.

Bretz F, W Maurer, W Brannath, and M Posch, 2009, A graphical approach to sequentially rejective multiple test procedures, Statistics in Medicine, 28: 586-604.

Bretz F, M Posch, E Glimm, F Klinglmueller, W Maurer, and K Rohmeyer, 2011, Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests, Biometrical Journal, 53(6): 894-913.

Committee For Proprietary Medicinal Products, 2002, Points to consider on multiplicity issues in clinical trials, London: The European Agency for the Evaluation of Medicinal Products, accessed December 1, 2020, http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf.

Dahlöf G, RB Devereux, SE Kjeldsen, S Julius, G Beevers, U de Faire, F Fyhrquist, H Ibsen, K Kristiansson, O Lederballe-Pedersen, LH Lindholm, MS Nieminen, P Omvik, S Oparil, H Wedel, and LIFE Study Group, 2002, Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol, Lancet, 359(9311): 995-1003.

Dmitrienko A, AC Tamhane, and BL Wiens, 2008, General Multistage Gatekeeping Procedures, Biometrical Journal, 50: 667-677.

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

Dmitrienko A, D'Agostino RB, Huque MF. Key multiplicity issues in clinical drug development. *Statistics in Medicine* 2013; **32**: 1079–1111.

Dmitrienko A, D'Agostino RB. Tutorial in Biostatistics: Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine* 2013; **32**(29): 5172-5218.

Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**: 800-802.

Hochberg Y and AC Tamhane, 1987, Multiple Comparison Procedures, New York (NY): John Wiley & Sons.

Holm SA, 1979, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics, 6(2): 65-70.

Hommel G, F Bretz, and W Maurer, 2011, Multiple hypotheses testing based on ordered p values — a historical survey with applications to medical research, Journal of Biopharmaceutical Statistics, 21(4): 595-609.

Hung HMJ and SJ Wang, 2010, Challenges to multiple testing in clinical trials, Biometrical Journal, 52(6): 747-756.

Huque MF, 2016, Validity of the Hochberg procedure revisited for clinical trial applications, Statistics in Medicine, 35(1):5-20.

Huque MF, M Alosh, and R Bhore, 2011, Addressing multiplicity issues of a composite endpoint and its components in clinical trials, Journal of Biopharmaceutical Statistics, 21: 610-634.

Huque MF, A Dmitrienko, and RB D'Agostino, 2013, Multiplicity issues in clinical trials with multiple objectives, Statistics in Biopharmaceutical Research, 5(4): 321-337.

Guidance for industry *E9 Statistical Principles for Clinical Trials* (September 1998)

Guidance for industry *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials* (October 2017)

Lubsen J and BA Kirwan, 2002, Combined endpoints: can we use them?, Statistics in Medicine, 21: 2959–2970.

Moye LA, 2000, Alpha calculus in clinical trials: considerations and commentary for the new millennium, Statistics in Medicine, 19:767-779.

Moye LA, 2003, Multiple Analyses in Clinical Trials, New York (NY): Springer-Verlag.

O'Neill RT, 1997, Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance, Controlled Clinical Trials, 18: 550-556.

Pocock SJ, CA Ariti, TJ Collier, and D Wang, 2012, The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities, European Heart Journal, 33: 176–182.

Sarkar S and CK Chang, 1997, Simes' method for multiple hypotheses testing with positively dependent test statistics, Journal of the American Statistical Association, 92: 1601-1608.

Westfall PH, RD Tobias, D Rom, RD Wolfinger, and Y Hochberg, 1999, Multiple Comparisons and Multiple Tests Using the SAS® System, Cary (NC): SAS Institute.

Westfall PH and SS Young, 1993, Resampling Based Multiple Testing: Examples and Methods for P-value Adjustment, New York (NY): Wiley-Interscience.

Wiens BL, 2003, A fixed sequence Bonferroni procedure for testing multiple endpoints, Pharmaceutical Statistics, 2: 211-215.

APPENDIX: STATISTICAL METHODS

This appendix presents some commonly used statistical methods and approaches for addressing multiplicity problems in controlled clinical trials that evaluate treatment effects on multiple endpoints. The methods listed in this appendix are not intended to be a comprehensive list of methods for controlling multiplicity; other approaches could be appropriate for specific situations. The choice of the method to use for a specific clinical trial will depend on the objectives and the design of the trial, as well as the knowledge of the drug being developed and the clinical setting. The method, however, should be decided upon prospectively. Because the considerations that go into the choice of multiplicity adjustment method can be complex and specific to individual product development programs, this guidance does not attempt to recommend any one method over another in most cases. Sponsors should consider the variety of methods available and in the prospective analysis plan select the most powerful method that is suitable for the design and objective of the study and maintains Type I error rate control.

1. The Bonferroni Method

The Bonferroni method is a single-step procedure that is commonly used, perhaps because of its simplicity and broad applicability. The drug is considered to have shown effects for each endpoint that succeeds on this test. The Holm and Hochberg methods (see below) are more powerful than the Bonferroni method for primary endpoints and are therefore preferable in many cases. However, sponsors might still wish to use the Bonferroni method for primary endpoints to maximize power for secondary endpoints or because the assumptions of the Hochberg method are not justified.

The most common form of the Bonferroni method divides the available total α (typically 0.05 two-sided) equally among the chosen endpoints. The method then concludes that a treatment effect is significant at the α level for each one of the m endpoints for which the endpoint's p-value is less than α/m. Thus, with two endpoints, the critical α for each endpoint is two-sided 0.025. The Bonferroni test can also be performed with different weights assigned to endpoints, with the sum of the relative weights equal to 1.0 (e.g., 0.4, 0.3, 0.2, and 0.1 for four endpoints). These weights should be prespecified in the design of the trial, taking into consideration the clinical importance of the endpoints, the likelihood of success, or other factors.

2. The Holm Procedure

The Holm procedure is a multi-step step-down procedure; it is useful for endpoints with any degree of correlation. It is less conservative than the Bonferroni method because a success with the smallest p-value (at the same endpoint-specific alpha as the Bonferroni method) allows other endpoints to be tested at larger endpoint-specific alpha levels than does the Bonferroni method. The algorithm for performing this test is as follows:

The endpoint p-values resulting from the completed study are first ordered from the smallest to the largest. Suppose that there are m endpoints to be tested and $p_{(1)}$ represents the smallest p-value, $p_{(2)}$ the next-smallest p-value, $p_{(3)}$ the third-smallest p-value, and so on.

i. The test begins by comparing the smallest p-value, $p_{(1)}$, to α/m, the same threshold used in the equally-weighted Bonferroni correction. If this $p_{(1)}$ is less than α/m, the treatment effect for the endpoint associated with this p-value is considered significant.

ii. The test then compares the next-smallest p-value, $p_{(2)}$, to an endpoint-specific alpha of the total alpha divided by the number of yet-untested endpoints (e.g., $\alpha/(m\text{-}1)$ for the second smallest p-value, a somewhat less conservative significance level). If $p_{(2)} < \alpha/(m\text{-}1)$, then the treatment effect for the endpoint associated with this $p_{(2)}$ is also considered significant.

iii. The test then compares the next ordered p-value, $p_{(3)}$, to $\alpha/(m\text{-}2)$, and so on until the last p-value (the largest p-value) is compared to α.

iv. The procedure stops, however, whenever a step yields a non-significant result. Once an ordered p-value is not significant, the remaining larger p-values are not evaluated and cannot be considered as statistically significant.

There is also a more general weighted version of Holm which allows unequal alpha allocation to the individual null hypotheses.

3. The Hochberg Procedure

The Hochberg procedure is a step-up testing procedure. It is more powerful than the Holm procedure (i.e., if a treatment effect is significant under Holm procedure it will be also significant under Hochberg procedure but not necessarily vice versa), but, unlike the Holm procedure, it controls the overall error rate only under certain assumptions. It compares the p-values to the same alpha critical values of α/m, $\alpha/(m\text{-}1)$, …, α, as the Holm procedure, but, in contrast to the Holm procedure, the Hochberg procedure is a step-up procedure. Instead of starting with the smallest p-value, the procedure starts with the largest p-value, which is compared to the largest endpoint-specific critical value (α). Also, essentially in the reverse of the Holm procedure, if the first test of hypothesis does not show statistical significance, testing proceeds to compare the second-largest p-value to the second-largest adjusted alpha value, $\alpha/2$. Sequential testing continues in this manner until a p-value for an endpoint is statistically significant, whereupon the Hochberg procedure provides a conclusion of statistically significant treatment effects for that endpoint and all endpoints with smaller p-values. For example, when the largest p-value is less than α, then the method concludes that there are significant treatment effects for all endpoints. In another situation, when the largest p-value is not less than α, but the second-largest p-value is less than $\alpha/2$, then the method concludes that treatment effects have been demonstrated for all endpoints except for the one associated with the largest p-value.

The Bonferroni and the Holm procedures are well known for being assumption-free. The methods can be applied without concern for the endpoint types, their statistical distributions, and the type of correlation structure. The Hochberg procedure, on the other hand, is not assumption-free in this way. The Hochberg procedure is known to provide adequate overall alpha-control for independent endpoint tests or for positively correlated dependent tests with standard test statistics in some cases (e.g., the test statistics are jointly bivariate normal). It is also a valid test procedure

when certain conditions are met. Various simulation experiments for the general case (e.g., for more than two endpoints with unequal correlation structures) indicate that the Hochberg procedure usually will, but is not guaranteed to, control the overall Type I error rate for positively correlated endpoints, but fails to do so for some negatively correlated tests (Sarkar et al. 1997, Huque 2016).

4. Prospective Alpha Allocation Scheme

The Prospective Alpha Allocation Scheme (PAAS) (Moye 2000) is a single-step method that has a slight advantage in power over the Bonferroni method. The method allows equal or unequal alpha allocations to all endpoints, but, as with the Bonferroni method, each specific endpoint receives a prospective allocation of a specific amount of the overall alpha. The alpha allocations are required to satisfy the equation:

$$(1 - \alpha_1)(1 - \alpha_2) \ldots (1 - \alpha_k) \ldots (1 - \alpha_m) = (1 - \alpha).$$

Each element in this equation, $(1 - \alpha_k)$, is the probability of correctly not rejecting the null hypothesis for the k^{th} endpoint, when it is tested at the allocated alpha α_k. This procedure is valid when the endpoints are independent or positively correlated, but the Type I error rate may be inflated when the endpoints are negatively correlated. This equation states the requirement that probability of correctly not rejecting all of the individual null hypotheses, calculated by multiplying each of the *m* probabilities together, should equal the selected goal (e.g., 0.95). The alpha allocation for any of the individual endpoint tests can be arbitrarily assigned, if desired, but the total group of allocations should always satisfy the above equation. In general, when arbitrary alpha allocations are made for some endpoints, at least the last endpoint's alpha should be calculated in order to satisfy the overall equation. As stated earlier, the Bonferroni method relies upon a similar constraint-defining equation, except that for the Bonferroni method the sum of all the individual alphas should equal the overall alpha.

5. The Fixed-Sequence Method

In many studies, testing of the endpoints can be ordered in a specified sequence, often ranking them by clinical relevance or likelihood of success. A fixed-sequence statistical testing procedure tests endpoints in a predefined order, all at the same significance level alpha (e.g., $\alpha = 0.05$), moving to the next endpoint only after a success on the previous endpoint. Such a testing procedure requires (1) prospective specification of the testing sequence and (2) no further testing once the sequence breaks; that is, further testing stops as soon as there is a failure of an endpoint in the sequence to show significance at level alpha (e.g., $\alpha = 0.05$).

The appeal of the fixed-sequence testing method is that it does not require any alpha adjustment of the individual tests. Its main drawback is that if a hypothesis in the sequence is not rejected, statistical significance cannot be achieved for the endpoints planned for the subsequent hypotheses, even if they have extremely small p-values. Suppose, for example, that in a study, the p-value for the first endpoint test in the sequence is $p = 0.59$, and the p-value for the second endpoint is $p = 0.001$; despite the apparent strong finding for the second endpoint, the result is not considered statistically significant. Ignoring the first endpoint's result recreates the

multiplicity problem and causes inflation of the overall Type I error rate. For this example, other methods of controlling Type I error such as the Bonferroni method, would have shown an effect for the second endpoint.

Thus, for the fixed-sequence method, carefully selecting the ordering of the tests of hypotheses is critical. A test early in the sequence that fails to show statistical significance will render the remainder of the endpoints not statistically significant. It is often not possible to determine a priori the best order for testing (Hung and Wang 2010), and there are other methods for addressing the multiplicity problem, which are described in the following subsections.

6. Resampling-Based, Multiple-Testing Procedures

When there is correlation among multiple endpoints, resampling (Westfall and Young 1993) is one general statistical approach that can provide more power than the methods described above to detect a true treatment effect while maintaining control of the overall Type I error rate, and the power increases as the correlation increases. With these methods, a distribution of the possible test-statistic values under the null hypothesis is generated based upon the observed data of the trial. This data-based distribution is then used to find the p-value of the observed study result instead of using a theoretical distribution of the test statistics (e.g., a normal distribution of Z-scores, or a t-distribution for t-scores) as with most other methods.

Resampling methods include the bootstrap and permutation approaches for multiple endpoints and require few, albeit important, assumptions about the true distribution of the endpoints. There are, however, some drawbacks to these methods. The important assumptions are generally difficult to verify, particularly for small study sample sizes. These methods, consequently, usually require large study sample sizes (particularly bootstrap methods) and often require simulations to ensure the data-based distribution of the test statistics from the limited trial data is applicable and to ensure adequate Type I error rate control. Inflation of the Type I error rate may occur, for example, if the shape of the data distribution is different between the treatment groups being compared.

7. Gatekeeping Testing Strategies

Gatekeeping procedures (e.g., Dmitrienko et al. 2008, Dmitrienko and D'Agostino 2013) address the problems of testing hierarchically ordered families of null hypotheses. Families usually correspond to primary and secondary objectives in a clinical trial (see section III.A.). Inferences in each family depend on the acceptance or rejection of null hypotheses in the earlier families consistent with logical relationships that may exist among the null hypotheses. The relationships usually reflect the relevant clinical considerations and are specified using a set of logical restrictions. Different types of logical gatekeeping constraints have been studied including serial gatekeeping, parallel gatekeeping and their generalization referred to as tree-structured gatekeeping.

A serial strategy can be applied, for example, in the scenario where the endpoints of the primary family are tested as co-primary endpoints (section III.C.). If all endpoints in the primary family are statistically significant at the alpha level (e.g., $\alpha = 0.05$), the endpoints in the second family

are examined. The endpoints in the second family can be tested at the overall alpha level by any prespecified acceptable method (e.g., Holm procedure, the fixed-sequence method, or others described in this appendix) that controls Type I error rate within the second family. If, however, at least one of the null hypotheses of the primary family fails to be rejected, the primary family criterion has not been met and the secondary endpoint family is not tested.

A parallel gatekeeping strategy is applied when the endpoints in the primary family are not all co-primary endpoints, and a separable testing method (e.g., Bonferroni method or Truncated Holm method) is specified for the primary family. In this strategy, the second endpoint family is examined when at least one of the endpoints in the first family has shown statistical significance.

Some multiplicity problems are multidimensional. One dimension may correspond to multiple endpoints, a second to multiple-dose groups (that have each of those endpoints tested), and yet another dimension to multiple hypotheses regarding an endpoint, such as non-inferiority and superiority tests (for each dose and each endpoint). The multiple sources of multiplicity create the potential for multiple pathways of testing the hypotheses. For example, if the goal of a study is to demonstrate non-inferiority as well as superiority, a single path of sequential tests is preferred. Suppose, however, that one wants to analyze a second endpoint for non-inferiority after the first endpoint is successfully shown to be non-inferior. The testing path now branches into two paths from this initial test (i.e., testing superiority for the first endpoint and non-inferiority for the second endpoint).

The multi-branched gatekeeping procedure allows for ordering the sequence of testing with the option of testing of more than one endpoint if a preceding test is successful. When there are multiple levels of this sequential hierarchy, and branching is applied at several of the steps, the possible paths of endpoint testing become a complex, multi-branched structure.

As a simple illustration (Figure A1), consider a clinical trial that compares a treatment to control on two primary endpoints (Endpoint 1 and Endpoint 2) to determine first whether the treatment is non-inferior to the control for at least one endpoint. If, for either of the two endpoints, the treatment is found non-inferior to the control, there is also a desire to test whether it is superior to control for that endpoint. The analytic plan for the trial thus sets the following logical restrictions:

i. Test endpoint two only after non-inferiority for endpoint one is first established.

ii. Test for superiority on an endpoint only after non-inferiority for that endpoint is first concluded.

The following diagram shows the decision structure of the test strategy. In this diagram, each block (or node) states the null hypothesis that it tests.

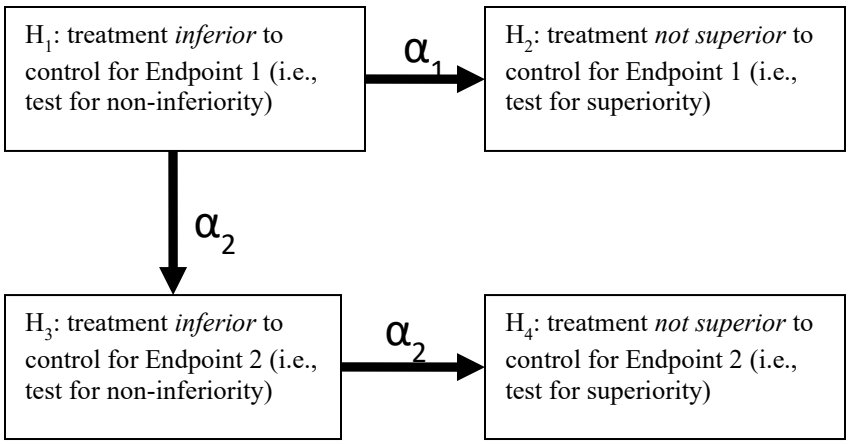

Figure A1: Example of a flow diagram for non-inferiority and superiority tests for endpoints one and two of a trial with logical restrictions, where $\alpha_1 + \alpha_2 = \alpha$. To test for superiority for Endpoint 1 and/or 2, one should first establish non-inferiority for that endpoint.

Thus, the above test strategy has a two-dimensional hierarchical structure, one dimension for the two different endpoints and the other for the non-inferiority and superiority tests, with the logical restrictions as stated above. Note that for this type of procedure, if multiple branches split off from a single node, the alpha should be split across the multiple branches.

8. Graphical Approaches Based on Sequentially Rejective Tests

The graphical approach (e.g., Bretz et al. 2009) is a means for developing and evaluating multiple analysis strategies for Bonferroni-based sequentially rejective methods. This approach illustrates differences in endpoint importance as well as the relationships among the endpoints by mapping onto a test strategy that ensures control of the Type I error rate and aids in creating and evaluating alternative test strategies.

Graphical displays of complex analysis strategies can aid in describing and assessing the proposed plan by displaying all the logical relationships among endpoint tests of hypotheses.

Basics of the Graphical Approach: Use of vertex (node) and path (order or direction)

In the graphical approach, the testing strategy is defined by a figure (graph) that shows each of the hypotheses (H_1, H_2,..., H_m) located at a vertex (or node, a junction of testing order paths). Each vertex (hypothesis) is allocated an initial amount of alpha, which this document defines as the endpoint-specific alpha (with the understanding that a test of an endpoint is associated with a test of a hypothesis, and vice versa). A key requirement is that the sum of all of the endpoint-specific alpha levels is equal to the total alpha level available for the study (the overall Type I error rate). At each step of the algorithm, endpoints are tested at the endpoint-specific significance levels using Bonferroni procedure.

Another feature of the figure (graph) is a set of directed edges. Each directed edge (or arrow) connects two hypotheses and is assigned a value between 0 and 1, called a weight for that edge and shown above the arrow, which indicates the fraction of the preserved alpha to be shifted

along that path to the receiving hypothesis, when the hypothesis at the tail end of the path is successful (i.e., is rejected). The sum of the weights across all the paths leaving a vertex should be 1.0, so that the entire preserved alpha is used in testing subsequent hypotheses. All study hypotheses that are intended to potentially provide firm conclusions of efficacy are shown in the graph.

Several examples of the graphical method follow to help illustrate the concept, construction, interpretation, and application of these diagrams.

Fixed-Sequence Method

The fixed-sequence testing strategy (appendix section 5.), shown in Figure A2, illustrates a simple case of the graphical method with three hypotheses. In this scheme, the endpoints (hypotheses) are ordered. Testing begins with the first endpoint at the full alpha level and continues through the sequence only until an endpoint is not statistically significant. This diagram shows that the endpoint-specific alpha levels associated with hypotheses H_1, H_2, and H_3 are set in the beginning as α, 0, and 0. For the fixed-sequence method, arrows represent the sequence of testing, and if the test is successful, the full alpha is shifted along to the next test. Consequently, if null hypothesis H_1 is successfully rejected, the endpoint-specific alpha level for H_2 becomes $0 + 1 \times \alpha = \alpha$, which allows testing of H_2 at level α. However, if the test of H_1 is unsuccessful, there is no pre-assigned non-zero alpha for H_2 to allow testing of H_2, so the testing stops.

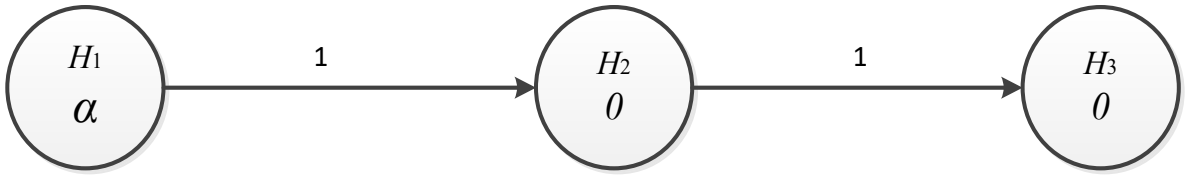

Figure A2: Graphical illustration of the fixed-sequence testing with three hypotheses.

Loop-Back Feature to Indicate Two-Way Potential for Retesting

Another valuable feature of the graphical method occurs when the available alpha level is split between two or more endpoints into endpoint-specific alpha levels; these diagrams illustrate the potential for loop-back passing of endpoint-specific alpha.

The Holm procedure (appendix section 2.) is a specific case of tests for two hypotheses with a loop-back feature where the graphical method enables a simple depiction of the procedure and its rationale. The Holm procedure directs that the first step is to test the smaller p-value at endpoint-specific alpha = $\alpha/2$ and, only if successful, proceed to test the larger p-value at the level α (e.g., 0.05). Because the Holm procedure splits alpha evenly in half, if the test of hypothesis with the smaller p-value was not significant, it is clear that the test with the larger p-value will also fail to be significant; performing that comparison is unnecessary. The diagram for the Holm procedure (Figure A3), shows two vertices and associated endpoint-specific alpha levels of $\alpha_1 = 0.025$ and $\alpha_2 = 0.025$, respectively, satisfying the requirement for total alpha = 0.05. The two arrows show that alpha might be passed along from H_1 to H_2, or H_2 to H_1. If the first test is successful, the endpoint-specific alpha of 0.025 is shifted entirely to the other hypothesis and added to the endpoint-specific alpha already allocated for that hypothesis to provide a net alpha of 0.05. Because either hypothesis might be tested first, the diagram shows a loop-back configuration.

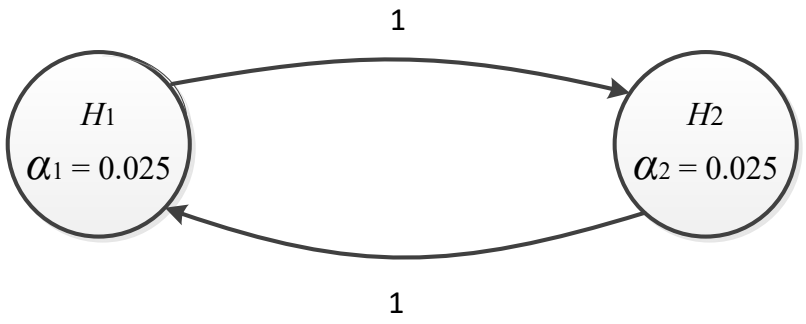

Figure A3: Graphical illustration of the Holm procedure with two hypotheses.

Testing on the diagram can start at any of the vertices that have non-zero alpha in the initial diagram, and all vertices with non-zero alpha can be tested until one is found for which the test is successful (i.e., the hypothesis is rejected). Then, the respective node is removed, and the alpha allocated to the rejected hypothesis propagates to other nodes following the arrows, as directed in the diagram. The final conclusions of which hypotheses were rejected and which were not will be the same irrespective of which vertex was inspected first. The graphical method enables complex alpha-splitting and branching of testing path features to be clearly identified as part of the analysis plan and correctly implemented.

Progressive Updating of the Diagram When Hypotheses Are Successfully Rejected

The graphical approach guides the hierarchical testing of multiple hypotheses through continual updating of the initial graph whenever a hypothesis is successfully rejected. The initial graph represents the full testing strategy (with all hypotheses). Each new graph shows the progression

of the testing strategy by eliminating hypotheses that have been rejected and retaining those yet to be tested or re-tested.

When there is a desire to consider analysis strategies with complex division of alpha, the graphical method and progressive updating of the diagram can aid in understanding the implication of the different strategies for a variety of different hypothetical scenarios. This progressive updating can aid in selecting which specific strategy to select for the final study statistical analysis plan.