Multiple Endpoints in Clinical Trials

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to http://www.regulations.gov. Submit written comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document contact (CDER) Scott Goldie at 301-796-2055 or (CBER) Office of Communication, Outreach, and Development, 800-835-4709 or 240-402-8010.

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)

[January 2017]
Clinical/Medical

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
http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

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
http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

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)

[January 2017]
Clinical/Medical

TABLE OF CONTENTS

Multiple Endpoints in Clinical Trials
Guidance for Industry[1]

This draft guidance, when finalized, will represent the current thinking of the Food and Drug Administration (FDA or Agency) on this topic.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 staff 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. As the number of endpoints analyzed in a single trial increases, the likelihood of making false conclusions about a drug's effects with respect to one or more of those endpoints becomes a concern if there is not appropriate adjustment for multiplicity. The purpose of this guidance is to describe various strategies for grouping and ordering endpoints for analysis and applying some well-recognized statistical methods for managing multiplicity within a study in order 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.

FDA's guidance for industry *E9 Statistical Principles for Clinical Trials* (International Council on Harmonisation E9 guidance, or "ICH E9")[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 at the Food and Drug Administration.

[2] The ICH E9 guidance is available on the FDA Drugs Web page under ICH – Efficacy. We update guidances periodically. To make sure you have the most recent version of a guidance, check the FDA Drugs Web page at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

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

Failure to account for multiplicity when there are several clinical endpoints evaluated in a study can lead to 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 and support drug approval; however, this issue is important throughout the drug development process.

A. Introduction to Study Endpoints

Efficacy endpoints are measures intended to reflect the effects of a drug. They include assessments of clinical events (e.g., mortality, stroke, pulmonary exacerbation, venous thromboembolism), patient symptoms (e.g., pain, dyspnea, depression), measures of function (e.g., ability to walk or exercise), or surrogates of these events or symptoms.

Because most diseases have more than one consequence, many trials are designed to examine the effect of a drug on more than one endpoint. In some cases, efficacy cannot be adequately established on the basis of a single endpoint. In other cases, an effect on any of several endpoints could be sufficient to support approval of a marketing application. When the rate of occurrence of a single event is expected to be low, it is common to combine several events (e.g., cardiovascular death, heart attack, and stroke) in a "composite event endpoint" where the occurrence of any of the events would constitute an "endpoint event."

When there are many endpoints prespecified in a clinical trial, they are usually classified into three families: primary, secondary, and exploratory.

- The set of primary endpoints consists of the outcome or outcomes (based on the drug's expected effects) that establish the effectiveness, and/or safety features, of the drug in order to support regulatory action. When there is more than one primary endpoint and success on any one alone could be considered sufficient to demonstrate the drug's effectiveness, the rate of falsely concluding the drug is effective is increased due to multiple comparisons (see section II.E).
- Secondary endpoints may be selected to demonstrate additional effects after success on the primary endpoint. For instance, a drug may demonstrate effectiveness on the primary endpoint of survival, after which the data regarding an effect on a secondary endpoint, such as functional status, would be tested. Secondary endpoints may also provide evidence that a particular mechanism underlies a demonstrated clinical effect (e.g., a drug for osteoporosis with fractures as the primary endpoint, and improved bone density as a secondary endpoint).
- All other endpoints are referred to as exploratory in this document (see section III.A).

Endpoints are frequently ordered by clinical importance, with the most important being designated as primary (e.g., mortality or irreversible morbidity). This is not always done, however, for a variety of reasons. The most common reasons not to order endpoints by clinical importance are if there are likely to be too few of the more clinically important endpoint events to provide adequate power for the study, or if the effect on a clinically less important endpoint is expected to be larger. In these cases, endpoints are often ordered by the likelihood of demonstrating an effect. For example, time-to-disease progression is often selected as the primary endpoint in oncology trials even though survival is almost always the most important endpoint; the reasons being that an effect on disease progression may be more readily demonstrable, may be detected earlier, and often has a larger effect size because the observed effect on survival can be diluted by subsequent treatment post-progression. Section III.A includes further discussion of the primary and secondary endpoint families. The determination of which endpoints are primary, secondary, or exploratory, regardless of the reasons for the determination, should always be made prospectively (see ICH E9).

Although this guidance focuses on endpoints intended to demonstrate effectiveness, a study that is designed specifically to assess safety outcomes may also have both primary and secondary endpoints, which would then be subject to the same multiplicity considerations described in this guidance.

B. Demonstrating the Study Objective of Effectiveness

A conclusion that a study has demonstrated an 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).[3] 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."[4] The clinical trial community has accepted an approach that finds a treatment effect to be established when a determination is made that the apparent treatment effect observed in a clinical trial is not likely to have occurred by chance. This is generally accomplished by placing a limit on the probability that the finding is the result of chance.

[3] Similarly, 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 biologics. 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).

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

The statistical approach commonly used to address the certainty/uncertainty in the assessment of a treatment effect on a chosen clinical endpoint is based on the *test of hypothesis*. This approach begins with stating the relevant hypotheses for each endpoint. In the simplest situation, two mutually exclusive hypotheses are specified for each endpoint in advance of conducting a clinical trial:

- One hypothesis, the *null hypothesis*, states that there is no treatment effect on the chosen clinical endpoint. The treatment effect is represented by a parameter, for example, T-C, the difference between the test group's average outcome measure (T) and that of the control group (C), or T/C, the ratio of response rates for the two groups. The null hypothesis is represented by the equation $T\text{-}C = 0$ or $T/C = 1$, stating that the true difference between the outcomes for the test group and the control group is zero or the risk ratio is 1 (i.e., there is no treatment effect).
- The other hypothesis is called the *alternative hypothesis* and posits that there is at least some treatment effect of the test drug, usually represented as $T\text{-}C > 0$ (or $T/C > 1$) for the alternative of interest (a beneficial effect of the drug).

The *test of hypothesis* determines whether (1) the trial results are consistent with the null hypothesis of no treatment effect or (2) the favorable result of the trial is so unlikely to have been obtained if the null hypothesis were true that the null hypothesis can be rejected and the alternative hypothesis, that there is a treatment effect, accepted.

Sometimes (e.g., in some vaccine trials), demonstration of an effect of at least some minimum size is considered essential for approval of a drug. In this case the null hypothesis might be modified to $T\text{-}C \leq m$ or $T/C \leq r$, where m or r is the smallest effect that could be accepted. Such modifications of the null hypothesis can have an impact on the sample size of a trial.

C. 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. 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. The probability of concluding that there was a difference between treatment groups due to the drug when, in fact, there was none, is called the Type I error probability or rate, denoted as alpha (α).

Type I error probabilities can apply to two-sided hypothesis tests, in which case they refer to the probability of concluding that there is a difference (beneficial or harmful) between the drug and control when there is no difference. Type I error probabilities can also apply to one-sided hypothesis tests, in which case they refer 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 alpha are 0.05 for two-sided tests and 0.025 for one-sided tests. In the case of one-sided tests, an alpha 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 percent, or 1 chance in 40 (represented as $p \leq 0.025$). In the case of two-sided tests, an alpha 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 percent, or 1 chance in 20 (represented as $p \leq 0.05$). Use of a two-sided test with an alpha of 0.05 generally also ensures that the probability of falsely concluding *benefit* when there is none is no more than approximately 2.5 percent (1 chance in 40). Use of either test therefore provides strong assurance against the possibility of a false-positive result (i.e., no more than 1 chance in 40) and a sound basis for regulatory decision-making, especially when substantiated by another study or other confirmatory evidence.[5]

For simplicity, this guidance discusses statistical testing of two-sided hypotheses at the 5 percent level, with the understanding that the one-sided alternative hypothesis of a beneficial drug effect is our focus, and the chance of a false positive conclusion is our primary concern. In most cases, sponsors can perform either two-sided or one-sided tests of hypothesis, at their discretion.

This discussion is focused on the study's final analysis. If interim analyses occur during a study, there should be a prospective plan to ensure that these additional analyses do not increase the chances of a false positive conclusion. When multiple endpoints are examined at an interim analysis, the appropriate adjustments can become complex; discussion of this issue is outside the scope of this guidance.

FDA's concern for controlling the Type I error probability is to minimize the chances of a false favorable conclusion for any of the primary or secondary endpoints, regardless of which and how many endpoints in the study have no effect (called strong control of the Type I error probability). Determining if strong control is achieved can be complicated when more than one endpoint is under consideration, any one of which could support a conclusion that the treatment has a beneficial effect. When there is more than one study endpoint, care must be taken to ensure that the evaluation of multiple hypotheses does not lead to inflation of the study's overall Type I error probability, called the study-wise Type I error probability, which is the chance of a false positive conclusion on any planned endpoint analysis.

The discussion of specific statistical methods for managing multiplicity in section IV illustrates that when some of the null hypotheses should be rejected but others should not be rejected, the control of the Type I error probability can become complex. The challenge that arises from testing multiple hypotheses associated with multiple endpoints in a study is to ensure that there is a proper accounting for all of the possible ways the endpoints of the study could produce false positive conclusions (see section II.E).

An essential element of Type I error rate control is the prospective specification of:

- all endpoints that will be tested and
- all data analyses that will be performed to test hypotheses about the prespecified endpoints.

For a multiple endpoints study, the analysis plan should describe how (or ways to determine how) the endpoints are tested, including the order of testing and the alpha level applied to each specific test.

[5] See the FDA guidance for industry *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*, available on the FDA Drugs guidance Web page under Clinical/Medical.

D. Relationship Between the Observed and True Treatment Effects

The statistical analysis associated with a hypothesis test produces three primary measures of interest:

- a point estimate,
- a confidence interval, and
- a p-value.

The effect of the treatment is typically presented as a point estimate (the observed T-C difference) that represents the most likely true effect. The confidence interval is usually two-sided and illustrates the range of true treatment effect values consistent with the data observed in the trial.

In addition to the point estimate of the treatment effect, it is important to consider the width of the confidence interval. The confidence interval provides a measure of the precision of the estimate of the treatment effect. The narrower the confidence interval, and the further away its lower bound is from the null hypothesis of no treatment effect (T-C = 0 or T/C = 1), the more confident we are of both the magnitude and existence of the treatment effect. Generally, the farther the lower bound of the confidence interval is from zero (or 1), the more persuasive (smaller) the p-value is and the lower the likelihood that the effectiveness finding was a chance occurrence.

There is usually a relationship between the test of a hypothesis and the confidence interval; each focuses on related but not identical questions:

- The test of a hypothesis focuses on whether or not there is an effect.
- The confidence interval focuses on the magnitude of the effect and the precision with which we know it.

The emphasis of this guidance is not on the confidence interval, but rather on the test of a hypothesis, where the issue is whether a treatment effect on a particular endpoint exists at all. Although confidence intervals are also critical to the interpretation of an effect when one exists, determining the confidence interval with some of the statistical methods for managing multiplicity described in section IV is complex. The primary goal of this guidance is to provide recommendations for designing studies that control the chances of erroneously concluding that a treatment is effective with respect to a particular endpoint. In some areas, however, confidence intervals are used to test hypotheses of the type described at the end of section II.B (e.g., T-C ≤ m). In these situations, it is critical to ensure that the confidence intervals appropriately reflect multiplicity of hypothesis tests.

E. Multiplicity

As described in section I.A, clinical trials often include more than one endpoint as an indicator of effectiveness. When a trial is designed so that more than one study endpoint or comparison (of treatment to control) could lead to a conclusion that effectiveness was established, testing each endpoint separately at $\alpha = 0.05$ will inflate the Type I error rate and overstate the statistical significance. The inflation of the Type I error rate can be quite substantial if there are many comparisons. Because this form of Type I error rate inflation is the result of multiple comparisons, it is termed a multiplicity problem.

In a clinical trial with a single endpoint tested at $\alpha = 0.05$, the probability of finding a difference between the treatment group and a control group by chance alone is at most 0.05 (a 5 percent chance). By contrast, if there are two independent endpoints, each tested at $\alpha = 0.05$, and if success on either endpoint by itself would lead to a conclusion of a drug effect, there is a multiplicity problem. For each endpoint individually, there is at most a 5 percent chance of finding a treatment effect when there is no effect on the endpoint, and the chance of erroneously finding a treatment effect on at least one of the endpoints (a false positive finding) is about 10 percent. More precisely, when the endpoints are independent, there is a 95 percent chance of correctly failing to detect an effect for each endpoint if there is no true effect for either endpoint. The chance of correctly failing to detect an effect on both endpoints together is thus 0.95 * 0.95, which equals 0.9025, and so the probability of falsely detecting an effect on at least one endpoint is 1 - 0.9025, which equals 0.0975. Without correction, the chance of making a Type I error for the study as a whole would be 0.1 and the study-wise Type I error rate is therefore not adequately controlled. The problem is exacerbated when more than two endpoints are considered. For three endpoints, the Type I error rate is 1 - (.95 * .95 * .95), which is about 14 percent. For ten endpoints, the Type I error rate is about 40 percent.

Even when a single outcome variable is being assessed, if the approach to evaluating the study data is to analyze multiple facets of that outcome (e.g., multiple dose groups, multiple time points, or multiple patient subgroups based on demographic or other characteristics) and regard the study as positive (i.e., conclude that the drug has been shown to produce a beneficial effect) if any one analysis is positive, the multiplicity of analyses causes inflation of the Type I error rate, thus increasing the probability of reaching a false conclusion about the effects of the drug. Similarly, application of more than one analytic approach to one endpoint introduces multiplicity by providing additional ways for the trial to be successful (to "win"). Examples include conducting both unadjusted and covariate-adjusted analyses, use of different analysis populations (intent-to-treat, completers, per protocol), use of different endpoint assessments (by investigator vs. a central endpoint assessment committee), and many others. By inflating Type I error, multiplicity produces uncertainty in interpretation of the study results such that the strength of a finding becomes unclear, and 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 Type I error rate at 5 percent. Among these are adjustments to the alpha level for determining that an individual endpoint test is positive, structuring the order in which the endpoints are tested, and others. These approaches are discussed in detail in section IV.

An important principle for controlling multiplicity is to prospectively specify all planned endpoints, time points, analysis populations, and analyses. Once these factors are specified, appropriate adjustments for multiple endpoints and analyses can be planned and applied, as needed. Changes in the analytic plan to perform additional analyses, however, 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. In the past, it was not uncommon, after the study was unblinded and analyzed, to see a variety of post hoc adjustments of design features (e.g., endpoints, analyses), usually plausible on their face, to attempt to elicit a positive study result from a failed study — a practice sometimes referred to as data-dredging. Although post hoc analyses of trials that fail on their prospectively specified endpoints may be useful for generating hypotheses for future testing, they do not yield definitive results. The results of such analyses can be biased because the choice of analyses can be influenced by a desire for success. The results also represent a multiplicity problem because there is no way to know how many different analyses were performed and there is no credible way to correct for the multiplicity of statistical analyses and control the Type I error rate. Consequently, post hoc analyses by themselves cannot establish effectiveness. Also, additional endpoints that have not been pre-specified or evaluated with adjustment for multiplicity when required cannot, in general, be used to demonstrate an effect of the drug, even in successful studies.

The multiplicity problem is also an issue in safety evaluations of controlled trials. With the exception of trials designed specifically to evaluate a particular safety outcome of interest, in typical safety assessments, there are often (1) no prior hypotheses, (2) many plausible analyses, (3) numerous safety findings that would be of concern, and (4) interest in both individual large studies and pooled study results. Moreover, it is difficult to discern what the analytic plan was and how it might have changed. There is no easy remedy for these issues, beyond recognition of the problems and a search for additional support that a finding is not a matter of chance. For example, it is more credible that there is a causal relationship between an observed adverse event and the drug, if the findings are consistent across studies; are predicted on the basis of recognized class effects, mechanism of drug action, or nonclinical studies; or are related to dose or exposure. The multiplicity problems for these types of safety analyses are outside the scope of this guidance.

The focus of this guidance is control of the Type I error rate for the planned primary and secondary endpoints of a clinical trial so that the major findings are well supported and the effects of the drug have been demonstrated. Once a trial is successful (demonstrates effectiveness or "wins" on the primary endpoint(s)), there are many other attributes of a drug's effects that may be described. Analyses that describe these other attributes of a drug can be informative and are often included in physician labeling.[6] Examples include: the time course of treatment effects;[7] the full distribution of responses amongst participants;[8] treatment effects on the components of a composite endpoint;[9] and treatment effects amongst subgroups.[10]

[6] FDA guidance for industry *Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products — Content and Format*, available on the FDA Drugs guidance Web page under Labeling.

[7] See, e.g., labeling for Pulmicort Flexhaler™ (budesonide) at http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/021949s006lbl.pdf.

[8] See, e.g., labeling for tetrabenazine at http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/206129Orig1s000lbl.pdf.

[9] See, e.g., labeling for COZAAR® (losartan potassium) at http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/020386s061lbl.pdf.

Nevertheless, it is important to understand that these descriptions with respect to additional attributes are not demonstrated additional effects of a drug unless the analyses were prespecified, and appropriate multiplicity adjustments were applied. Therefore, presenting p-values from descriptive analyses (that is, from analyses that were not prespecified and for which appropriate multiplicity adjustments were not applied) is inappropriate because doing so would imply a statistically rigorous conclusion and convey a level of certainty about the effects that is not supported by that trial. Descriptive analyses are not the subject of this guidance and are not addressed in detail.

In the following sections, the issues of multiple endpoints and methods to address them are illustrated with examples of different study endpoints. Both the issues and methods that apply to multiple endpoints also apply to other sources of multiplicity, including multiple doses, time points, or study population subgroups.

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 essential to establish effectiveness for approval are called *primary endpoints*. *Secondary endpoints* may be used to support the primary endpoint(s) and/or demonstrate additional effects. The third category in the hierarchy includes all other endpoints, which are referred to as exploratory. *Exploratory endpoints* may include clinically important events that are expected to occur too infrequently to show a treatment effect or endpoints that for other reasons are thought to be less likely to show an effect but are included to explore new hypotheses. Each category in the hierarchy may contain a single endpoint or a family of endpoints.

1. Primary Endpoint Family

The endpoint(s) that will be the basis for concluding that the study met its objective (i.e., the study "wins") is designated the primary endpoint or primary endpoint family. When there is a single pre-specified primary endpoint, there are no multiple endpoint-related multiplicity issues in the determination that the study achieved its objective; however, there could still be multiplicity issues for demonstration of effects on secondary endpoints.

Multiple primary endpoints occur in three ways, further described in section III.C. The first is when there are multiple primary endpoints corresponding to multiple chances to "win," and in

[10] See, e.g., labeling for BRILINTA® (ticagrelor) at http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/022433s017lbl.pdf.

this case, failure to adjust for multiplicity can lead to a false conclusion that the drug is effective. The second is where determination of effectiveness depends on success on all of two or more primary endpoints. In this setting, there are no multiple endpoint-related multiplicity issues, and therefore, no concern with Type I error rate inflation, but there is a concern with Type II error rate inflation (See Section III.B). 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 (section III.A.2). A comprehensive examination of the drug's effects earlier in development might aid in the selection of a sensitive and informative measure of the drug's effect and allow use of a single primary endpoint for the confirmatory trial.

2. Secondary Endpoint Family

The collection of all secondary endpoints is called the secondary endpoint family. Secondary endpoints are those that may provide supportive information about a drug's effect on the primary endpoint or demonstrate additional effects on the disease or condition. Secondary endpoints might include a pharmacodynamic effect that would not be considered an acceptable primary efficacy endpoint but is closely related to the primary endpoint, (e.g., an effect consistent with the drug's purported mechanism of action). 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). In all cases, when an effect on the primary endpoint is shown, the secondary endpoints can be examined and may contribute important supportive information about a drug's effectiveness.

Positive results on the secondary endpoints can be interpreted *only* if there is first a demonstration of a treatment effect on the primary endpoint family. The Type I error rate should be controlled for the entire trial, defined in section II.C as strong control. This includes controlling the Type I error rate within and between the primary and secondary endpoint families. Moreover, the Type I error rate should be controlled for any preplanned analysis of pooled results across studies; pooled analyses are rarely conducted for the planned primary endpoint, but are sometimes used to assess lower frequency events, such as cardiovascular deaths, where the individual trials used a composite endpoint, such as death plus hospitalization. Statistical testing strategies to accomplish this are discussed in section IV. Control of the Type I error rate for all endpoints depends upon the prospective designation of all primary and secondary endpoints. Generally, the endpoints and analytical plan should be provided at the time the trial protocol is finalized. The statistical analysis plan should not be changed after unmasking of treatment assignments, including unmasking for any interim analyses.

Because study sample size is often determined based only on the amount of information needed to adequately assess the primary hypothesis, many studies lack sufficient power to demonstrate effects on secondary endpoints. If success on the secondary endpoints is important, the secondary endpoints should be considered when determining study design (e.g., sample size).

An example of a secondary endpoint used to further characterize the drug's effect is a measurement of the primary outcome variable at 30 days in a trial whose primary endpoint is the same outcome measured at 6 months. Another example is a secondary endpoint of the percentage of patients whose symptoms are "very improved," when the primary endpoint is the percentage of patients with any amount of improvement for the same symptoms. Adjustment for multiplicity is necessary to demonstrate these additional effects.

It is recommended that the list of secondary endpoints be short, because the chance of demonstrating an effect on any secondary endpoint after appropriate correction for multiplicity becomes increasingly small as the number of endpoints increases. Endpoints intended to serve the purpose of hypothesis generation should not be included in the secondary endpoint family. These should be considered exploratory endpoints.

B. Type II Error Rate and Multiple Endpoints

One of the greatest concerns in the design of clinical trials intended to support drug approval is inflation of the Type I error rate, because it can lead to an erroneous conclusion that a drug is effective. FDA is also concerned with the risk of Type II error, which is failing to show an effect of a drug where there actually is one. The intended level of risk of a Type II error is usually denoted by the symbol beta (β). The study's likelihood of avoiding Type II error ($1-\beta$), if the drug actually has the specified effect, is called study power. The desired power is an important factor in determining the sample size.

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 is in fact present. In addition to the treatment effect, the optimal sample size of a study is influenced by the variability of the endpoint and the alpha level specified for the test of hypothesis for that endpoint. Investigators should consider these factors for all of the endpoints for which the study is intended to be well powered.

Many of the statistical adjustment methods to control the Type I error rate for multiplicity discussed in section IV decrease study power because they lower the alpha level used for each of the individual endpoints' test of hypothesis, making it more difficult to achieve statistical significance. Increasing the sample size appropriately can overcome this decrease in power. In general, the greater the number of endpoints (analyses), the greater the statistical adjustment that is needed and the greater the increase in the sample size of the trial necessary to maintain power for all individual endpoints. This decrease in study power (i.e., increased Type II error rate) from multiplicity is often a practical limiting factor in choosing the number of endpoints designated for a trial as indicators of success without requiring an excessive sample size.

Some of the methods discussed in section IV to manage multiplicity are complex and may, for example, call for the alpha level for any particular test of hypothesis to be determined by the actual study endpoint results and the resulting sequence of hypothesis testing. In some cases, sponsors may wish to have the study well powered for one or two secondary endpoints in addition to the primary endpoint family, further adding to the complexity. Determination of an appropriate study sample size to ensure that the study is appropriately powered can be difficult in these cases, and often will be dependent upon computer simulations rather than an analytic formula, which can be used for simpler situations.

The use of two or more endpoints for which demonstration of an effect on each is needed to support regulatory approval (called co-primary endpoints; see section III.C.1 below) increases the Type II error rate and decreases study power. If, for example, the study sample size is selected to provide 80 percent power to show success on each of two endpoints (i.e., Type II error rate is 20 percent for each), and the endpoints are entirely independent, the power to show success on both will be just 64 percent (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 percent. The study power could, of course, be restored by increasing the sample size. Multiplicity and Type I error rate inflation are not a concern with co-primary endpoints, as there is only one way to succeed.

The loss of power may not be so severe when the endpoints are correlated (i.e., not fully independent). With positive correlation, there is an increased chance that a second endpoint will demonstrate the treatment effect if one endpoint is successful, potentially increasing study power well above the 64 percent estimate. Moreover, the individual endpoints usually do not all have the same power-influencing characteristics because the effect size and variability estimates may be different for the different endpoints. If the study is designed so that a test of the endpoint upon which it is most difficult to demonstrate an effect has 80 percent power, the other endpoints may have power in excess of 80 percent to show the expected effect. In that case, the overall study power, even if the endpoints were fully independent, will also be higher than if all endpoints were equally powered. Nonetheless, when considering use of co-primary endpoints in a study, it should be recognized that use of more than two can markedly reduce study power.

C. Types of Multiple Endpoints

Multiple endpoints may be needed when determining that the drug confers a clinical benefit depends on more than one disease aspect or outcome being affected. Multiple endpoints may also be used when (1) there are several important aspects of a disease or several ways to assess an important aspect, (2) there is no consensus about which one will best serve the study purposes, and (3) an effect on any one will be sufficient as evidence of effectiveness to support approval. In some cases, multiple aspects of a disease may appropriately be combined into a single endpoint, but subsequent analysis of the components is generally important for an adequate understanding of the drug's effect. These circumstances when multiple endpoints are encountered are discussed below.

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

The primary endpoint for determining that a drug is effective should encompass one or more of the important features of a disorder and should be clinically meaningful. There are two types of circumstances when no single endpoint adequately serves this purpose.

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 it is necessary to demonstrate an effect on each of the endpoints to conclude that a drug is effective.

Therapies for the treatment of migraine headaches illustrate this circumstance. Although pain is the most prominent feature, migraine headaches are also often characterized by the presence of photophobia, phonophobia, and nausea, all of which are clinically important. Which of the three is most clinically important varies among patients. A recent approach to studying treatments is to consider a drug effective for migraines only if pain and an individually-specified most bothersome second feature are both shown to be improved by the drug treatment.

A second kind of circumstance in which a demonstration of an effect on two endpoints is needed is when there is a single identified critical feature of the disorder, but uncertainty as to whether an effect on the endpoint alone is clinically meaningful. In these cases, two endpoints are often used. One endpoint is specific for the disease feature intended to be affected by the drug but not readily interpretable as to the clinical meaning, and the second endpoint is clinically interpretable but may be less specific for the intended action of the test drug. A demonstration of effectiveness is dependent upon both endpoints showing a drug effect. One endpoint ensures the effect occurs on the core disease feature, and the other ensures that the effect is clinically meaningful.

An example illustrating this second circumstance is development of drugs for treatment of the symptoms of Alzheimer's disease. Drugs for Alzheimer's disease have generally been expected to show an effect on both the defining feature of the disease, decreased cognitive function, and on some measure of the clinical impact of that effect. Because there is no single endpoint able to provide convincing evidence of both, co-primary endpoints are used. One primary endpoint is the effect on a measure of cognition in Alzheimer's disease (e.g., the Alzheimer's Disease Assessment Scale-Cognitive Component), and the second is the effect on a clinically interpretable measure of function, such as a clinician's global assessment or an Activities of Daily Living Assessment.

Trials of combination vaccines are another 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 II.E, multiplicity problems occur when there is more than one way to determine that the study is a success. When using co-primary endpoints, however, there is only one result that is considered a study success, namely, that all of the separate endpoints are statistically significant. Therefore, testing all of the individual endpoints at the 0.05 level does not cause inflation of the Type I error rate; rather, the impact of co-primary endpoint testing is to increase the Type II error rate. The size of this increase will depend on the correlation of the co-primary endpoints. 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 relaxed (e.g., testing at an alpha of 0.06 or 0.07) to accommodate the loss in statistical power arising from the need to show an effect on both endpoints. Relaxation of alpha is generally not acceptable because doing so would undermine the assurance of an effect on each disease aspect considered essential to showing 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 may 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 drug 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 to be 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; section IV is devoted to describing and discussing some of these approaches.

It should be noted that failure to demonstrate an effect on any one of the individual prespecified primary endpoints does not preclude making valid conclusions with respect to the other prespecified primary endpoints. From a regulatory perspective, the results for all of the prespecified primary endpoints, both positive and negative, are considered in the overall assessment of risks and benefits.

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 may be appropriate to combine those clinical outcomes into a single variable. This is called a "composite endpoint," where an endpoint is defined as the occurrence or realization in a patient of any one of the specified components. When the components correspond to distinct events, composite endpoints are often assessed as the time to first occurrence of any one of the components, but in diseases where a patient might have more than one event, it also may be possible to analyze total endpoint events (see section III.D.1). A single statistical test is performed on the composite endpoint; consequently, no multiplicity problem occurs and no statistical adjustment is needed.

An important reason for using a composite endpoint is that the incidence rate 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 morbid, clinically important but uncommon events (e.g., use of an anti-platelet drug in patients with coronary artery disease to prevent myocardial infarction, stroke, and death).

The choice of the components of a composite endpoint should be made carefully. Because the occurrence of any one of the individual components is considered to be an endpoint event, each of the components is of equal importance in the analysis of the composite. The treatment effect on the composite 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 drug effect is chiefly on the least important event. Furthermore, it is possible that a component with greater importance may appear to 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 successful, careful examination of the data may 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.

4. Other Multi-Component Endpoints

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

When the components are ordered categorical or continuous numeric scales, one way of forming an overall rating is to use the sum or average across the individual domain scores. Study hypotheses are then tested by comparing the overall mean values between groups. Examples of this type are the Positive and Negative Syndrome Scale (PANSS) in schizophrenia research; the Toronto Western Spasmodic Torticollis Rating Scale for evaluating cervical dystonia; the Hamilton Rating Scale for Depression (HAM-D); the Brief Psychiatric Rating Scale; and many patient-reported outcomes (PROs).

Alternatively, a multi-component endpoint may be a dichotomous (event) endpoint corresponding to an individual patient achieving specified criteria on each of the multiple components. This dichotomous form of a multi-component endpoint might be preferred over multiple independent endpoints in conditions where assuring individual patients have benefit on all of several disease features is important. For example, the FDA guidance for industry *Considerations for Allogeneic Pancreatic Islet Cell Products* recommends that the primary endpoint in clinical trials of allogeneic pancreatic islet cells for Type 1 diabetes mellitus be a composite in which patients are considered responders only if they meet two dichotomous response criteria: normal range of HbA1c and elimination of hypoglycemia. In contrast, when separate endpoints are analyzed as co-primary endpoints (i.e., all of the several identified disease aspects are required to show an effect), the study provides evidence that the drug affects all of the endpoints on a group-wise basis, but does not ensure an increase in the number of individual patients for whom all endpoints are favorably affected.

More complex endpoint formulations may be appropriate when there are several different features of a disease that are important, but not all features must be positively affected for a patient to be regarded as receiving benefit. For example, a positive response for an individual patient might be defined as improvement in one or two specific required aspects of a disease along with improvement in at least one, but not all, identified additional disease features, as in the American College of Rheumatology (ACR) scoring system for rheumatoid arthritis. The ACR20 criteria for defining a response to treatment are a 20 percent improvement in two specific disease features (tender joints and swollen joints) and a 20 percent improvement in at least three of five additional features (pain, acute phase reactants, global assessment by patient or physician, or disability). Generally, these types of endpoints are very disease-specific, and clinical research on the particular disease and its manifestations guides the development of such defined, complex combinations of assessments. These combinations, despite incorporating multiple different features of the disease, provide a single primary endpoint for evaluating efficacy and do not raise multiplicity concerns.

The use of within-patient multi-component endpoints can be efficient if the treatment effects on the different components are generally concordant. Study power can be adversely affected, however, if there is limited correlation among the endpoints. Although multi-component endpoints can provide some gains in efficiency compared to co-primary endpoints, the appropriateness of a particular within-patient multi-component endpoint is generally determined by clinical, rather than statistical, considerations. Formal statistical analyses of these components without prespecification and adjustment for multiplicity, however, may lead to a false conclusion about the effects of the drug with respect to each individual component, as discussed in 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 may 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 clinical endpoint, if the effect of the drug on the composite primary endpoint is demonstrated. There have been situations, however, where the effect on the primary endpoint was not found to be statistically significant, but there did appear to be an effect on mortality or major morbidity. In the absence of a demonstrated treatment effect on the primary endpoint, secondary endpoints cannot be assessed statistically, but the suggestion of a favorable result on a major outcome such as mortality may be difficult to ignore.

One approach to avoid this situation would be to designate the mortality or morbidity endpoint as another primary endpoint, and apply one of the statistical methods of section IV with unequal splitting of the alpha. In this way, the endpoint can be validly tested, and should the effect be large, it will provide evidence of efficacy. Depending upon how alpha is allocated, the increase in sample size to maintain study power may only be modest.

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

1. Evaluating the Components 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 is 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 always be included in study reports. These analyses will not alter a conclusion about the statistical significance of the composite primary endpoint and are considered descriptive analyses, not tests of hypotheses. If there is, however, 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.

In analyzing the contribution of each component of a composite endpoint, there are two approaches that differ in how patients who experience more than one of the event-types are considered.

- One approach considers only the initial event in each patient. This method displays the incidence of each type of component event only when it was the first event for a patient. The sum of the first events across all categories will equal the total events for the composite endpoint.
- The other approach considers the events of each type in each patient. With this method, each of the components can also be treated as a distinct endpoint, irrespective of the order of occurrence, giving the numbers of patients who ever experienced an event of each type. In this case, each patient can be included in the event counts for more than one component, and the sum of events on all component types will be greater than the total number of composite events using only the first events.

An example to illustrate these approaches is the RENAAL trial, a study of the ability of losartan to delay development of diabetic nephropathy.[11] The primary endpoint was a composite endpoint of time to first occurrence of any one of three components: doubling of serum creatinine, progression to end-stage renal disease (ESRD), or death. Table 1 shows the crude incidence composite endpoint: there were 327 composite events in the losartan arm and 359 in the placebo arm, which led to a statistically-significant difference in the time-to-event analysis. The number of patients with an endpoint event at the end of study is tabulated in two ways. First, the decomposition of the composite endpoint events shows only events that were the first event for a patient. Thus, in the losartan arm, 162 patients had doubling of serum creatinine as a first event, 64 had ESRD, and 101 death. The total is 327, the same number as for the overall composite event, because only first events are counted. Table 1 includes the hazard ratio, confidence interval, and p-value for the primary composite endpoint. The confidence intervals and p-values are not given for the individual elements of the composite endpoint, because they were not designated as secondary endpoints and adequate control for multiplicity was not specified to support their assessment.

Table 1. Decomposition of Endpoint Events in RENAAL*

Endpoint	Losartan (N=751)	Placebo (N=762)	Hazard ratio± (95% CI)	p-value
Primary endpoint				
Doubling of serum creatinine, ESRD, or death	327	359	0.84 (0.72, 0.97)	0.022
Decomposition of the primary endpoint				
Doubling of serum Creatinine	162	198	0.75	
ESRD	64	65	0.93	
Death	101	96	0.98	
Any occurrence of individual components				
Doubling of serum Creatinine	162	198	0.75 (0.61, 0.92)	
ESRD	147	194	0.71 (0.57, 0.89)	
Death	158	155	1.02 (0.81, 1.27)	

*Excerpted from FDA/CDER/DBI Statistical Review at (http://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/20-386s028_Cozaar.cfm).
ESRD = end-stage renal disease; ±Hazard ratio from Cox proportional hazards time-to-event analysis.

The second analysis showing the results for any occurrence of individual components is quite different from the first-event-only decomposition analysis. There are now more total events, because some patients experience more than one event type and these patients are included in both component-event counts. In this example, ESRD events at any time yield a hazard ratio of 0.71, which is markedly different from that obtained for ESRD in the first-event only analysis,

[11] RENAAL: The Reduction of Endpoints in NIDDM with the Angiotensin II Antagonist Losartan Study.

0.93. Thus, the decomposition analysis limited to first events does not fully characterize the effect of losartan on ESRD.

The analysis of any occurrence of an event type, however, can be complicated by the issue known broadly in statistics as competing risks. This is the phenomenon wherein occurrence of certain endpoints can make it impossible to observe other events in the same patient. For example, in the RENAAL trial, patients whose first event was death could never be observed to have doubling of serum creatinine. If one study group had higher early mortality, it could appear to have a favorable profile with respect to other endpoint events simply because fewer patients survived, diminishing the number of patients at risk for the other types of events.

Study design and patient management issues can also complicate interpretation of the decomposition analyses. For example, in some trials, experiencing any endpoint event is cause to remove a patient from study therapy and to initiate treatment with alternative agents, including the possibility of receiving another treatment in the trial. Such a change in therapy obscures the relationship between the initial study therapy and the occurrence of subsequent events, so that only the analysis of first event will be useful. The complexities of interpretation of the decomposition analyses are important to consider when planning studies with a composite endpoint.

2. Reporting and Interpreting the Individual Component Results of a Composite Endpoint

The different components of a composite endpoint are selected because they are all clinically important; however, because each one is not necessarily equally affected by the drug, it is relevant and important to examine the effects of the drug on the individual components as well as on the overall endpoint. Presenting only data on the composite might imply meaningful treatment effects on all of the individual components, when a composite effect may in fact be established with little or no evidence of effect on some of the individual components. On the other hand, showing the results of the analysis for each of the individual components may imply an effect on an individual component when an appropriate statistical analysis would not support that conclusion. Thus, it is important to present descriptive analyses of between-group differences for the components in a way that does not overstate the conclusions.

It is common for one component of a composite endpoint to overly influence the treatment effect, but even if that is not so, and all components contribute, the inclusion of a particular component in a composite does not usually support an independent conclusion of efficacy on that component. FDA's guidance for industry *Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products — Content and Format*[12] calls for presentation in labeling of the components of a composite endpoint but without a statistical analysis of the separate components unless the components were prespecified as separate endpoints and assessed with a prospectively defined hypothesis and statistical analysis plan. In such a case, the statistical analysis will usually consider all events of each type, not just first-occurring events (as illustrated in Table 1 above). Only findings on prespecified endpoints that are statistically

[12] Available on the FDA Drugs Guidance Web page under Labeling.

significant, with adjustment for multiplicity, are considered demonstrated effects of a drug. All other findings are considered descriptive and would require further study to demonstrate that they are true effects of the drug. For example, a composite endpoint that includes mortality as a component provides little information about effects on mortality if there are few deaths, and presentations can make that clear by showing the actual numbers of deaths. Therefore, clear presentation of the results of the components of a composite is essential to describe where the drug's effect occurs. For example, the LIFE trial comparing losartan and atenolol in people with hypertension showed a clear, statistically-significant advantage of losartan on the composite endpoint of death, nonfatal myocardial infarction, or stroke, but this appeared to be related to an effect on fatal and nonfatal stroke, with no advantage on the incidence of acute myocardial infarction or cardiovascular death.[13]

To demonstrate an effect on a specific component or components of a composite endpoint, the component or components should be included prospectively as a secondary endpoint for the study or possibly as an additional primary endpoint (see section III.C.5), with appropriate Type I error rate control. If control of the Type I error rate is ensured with respect to the individual component or components, in addition to control for the composite, a trial will be potentially able to support conclusions regarding drug effects on the individual component or components as well as the composite.

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

As with composite endpoints, understanding which components of a within-patient multi-component endpoint (e.g., symptom rating scale such as HAM-D) have contributed most to the overall statistical significance could be important to correctly understanding the clinical effects of the drug. Consequently, a descriptive analysis of the study results on the individual components (or, in some cases, groups of similar components) may be considered but, as stated previously, if undertaken, should be presented in a way that does not overstate the conclusions. Unlike the composite endpoint used for outcome studies, where each component usually has clear clinical importance (death, acute myocardial infarction, stroke, hospitalization), the clinical importance of the components of these patient assessments may be less clear. Thus, for many of these multi-component endpoints, the overall score is regarded as comprehensive and clinically interpretable. The individual components of the scales, however, may not be independently clinically interpretable. Although some rating scales have been developed with broad multicomponent domains to allow the domains to be interpretable subsets of the overall scale, the individual domain and subscale scores generally are not prespecified for hypothesis testing. Prespecification of subscale scores with appropriate multiplicity control is required if it is thought to be important to demonstrate an effect of a drug on one or more of these subscale scores in addition to the overall multi-component endpoint.

Analyses of specific component item(s) of a symptom rating scale as explicit endpoints in the primary or secondary endpoint families may be reasonable, contingent on being clinically interpretable, in two cases:

---

[13] LIFE: The Losartan Intervention For Endpoint reduction in hypertension study.

(1) where earlier trials have suggested targeted efficacy of a drug on one or a small number of specific symptoms, or
(2) where the specific symptom measured by the item is considered to be of substantial inherent clinical importance.

An example of the first type is a novel agent for rheumatoid arthritis that was found in a controlled phase 2 trial to be particularly effective in lessening patients' pain. In this example, a sponsor might wish to test this hypothesis using a pain scale as a secondary endpoint in a trial where improvement meeting ACR20 criteria, which include pain as a component, is the primary endpoint. An example of the second type of component analysis might be found in trials of anti-psychotic drugs, in which positive and negative symptoms are domains collected in the Positive and Negative Syndrome Scale (PANSS) and often analyzed separately in addition to the overall scale. Interpretation of analyses of any subscale domain, however, is dependent on that subscale domain having been previously evaluated and determined to be valid as a stand-alone clinical measure. As described above (see section III.C), control of the Type I error rate will still be necessary for both the primary and secondary endpoint families.

IV. STATISTICAL METHODS

A variety of situations in which multiplicity arises have been discussed in sections II and III. Statistical methods provide acceptable ways to correct for multiplicity and control the Type I error rate for many of them. Standard statistical methods are available, for example:

- to examine treatment effects for multiple endpoints where success on any one endpoint would be acceptable, and
- to allow sequential testing where success on one endpoint permits analysis of additional endpoints.

This section describes methods that are commonly used for handling multiplicity problems in controlled clinical trials that examine treatment effects on multiple endpoints.

A. Type I Error Rate for a Family of Endpoints and Conclusions on Individual Endpoints

When there is a family of endpoints (discussed in sections II.A and III.A), the Type I error rate commonly used for the group of study endpoints is called the family-wise Type I error rate (FWER) or the overall Type I error rate for the family. The FWER is 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 within the family. This error rate is typically held to 0.05 (0.025 for one-sided tests). The statistical methods discussed in section IV.C maintain control of the FWER for finding significant treatment effects for study endpoints individually, thereby permitting an individual effectiveness conclusion on each endpoint.

There are also other statistical analysis methods, often called global procedures, that control the FWER with regard to erroneously concluding that there is a treatment effect on some endpoint (one or more) when there is no such effect on any endpoint. These methods allow a conclusion of treatment effectiveness in the global sense, but do not support reaching conclusions on the individual endpoints within the family. These methods are generally not encouraged when study designs and methods that test the endpoints individually are feasible; therefore, these global procedures are not described in this guidance.

Because composite and other multi-component endpoints (see sections III.C.3 and III.C.4) are constructed as a single endpoint, when they are part of an endpoint family, the methods described in section IV.C can be applied to them.

B. When the Type I Error Rate Is Not Inflated or When the Multiplicity Problem Is Addressed Without Statistical Adjustment or by Other Methods

This section identifies two situations involving multiple endpoints where inflation of the Type I error rate is avoided so that adjustments for multiplicity are not needed. These situations assume that the trial has no interim analysis or mid-course design modifications.

1. Clinically Relevant Benefits Required for All Specified Primary Endpoints — the Case of "Co-Primary" Endpoints[14]

As discussed in detail in section III.C, when multiple primary endpoints are tested and success in the study depends on success on all endpoints (i.e., they are co-primary endpoints), no multiplicity adjustment is necessary because there is no opportunity to select the most favorable result from among several endpoints. The impact of multiplicity in these situations is to increase the Type II error rate (section III.B).

2. Use of Multiple Analyses Methods for a Single Endpoint after Success on the Prespecified Primary Analysis Method

For many trials there are a range of plausible, closely related analyses of an individual endpoint. For example, the primary analysis of an outcome trial could adjust for certain covariates, make a different choice of covariates, make no covariate adjustment, be conducted on the intent-to-treat (ITT) population or various modified populations, or use various hypothesis testing methods. Accepting any one of these multiple analyses, when successful, as a basis for a conclusion that there is a treatment effect would increase the study Type I error rate, but it is difficult to estimate the increase in error rate because the results of these different analyses are likely to be similar and it is unclear how many choices could have been made. As with other multiplicity problems, prospective specification of the analysis method will generally eliminate the concern about a biased (result-driven) choice.

Once the effect has been clearly demonstrated based on the prespecified primary analysis, alternative analyses of the primary endpoint may be needed to correctly interpret the study's results. Additional analyses of the primary endpoint may be needed to gain a better

[14] Section 505(d) of the FD&C Act.

understanding of the observed treatment effect (e.g., to use a less conservative analysis to better estimate the effect size). In other cases, multiple related analyses are used to assess the sensitivity of the results to the important underlying assumptions of the prespecified analysis method. For example, sensitivity analyses may be needed to determine the impact of missing data on the primary analysis results, when the primary analysis method relies on unverifiable assumptions about those missing data. Note that these additional analyses do not demonstrate any new effects of the drug; rather, they clarify the effect already demonstrated by the primary analysis of a successful study.

C. Common Statistical Methods for Addressing Multiple Endpoint-Related Multiplicity Problems

This section presents some common statistical methods and approaches for addressing multiplicity problems in controlled clinical trials that evaluate treatment effects on multiple endpoints. 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 disorder. 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. There are, for example, a small number of situations in which one method is unambiguously more powerful than another without inflating the Type I error rate beyond the nominal level (e.g., the Holm method is more powerful than the Bonferroni method for primary endpoints). These situations are noted below.

The methods presented here are general, and the discussions and hypothetical examples have been generally limited to two-arm trials that examine treatment versus control differences on multiple endpoints. Similar considerations may apply to other kinds of multiplicity, such as in assessing treatment effects at different time points, or at different doses. Although the following discussions are oriented to the general reader, application of many of these methods can be technically complex and should be used relying on statistical expertise. Consequently, when a multiple endpoints problem arises in designing a clinical trial and one or more of these methods are to be considered, consultation with knowledgeable experts is important.

Statistical methods for addressing multiplicity issues are broadly classified into two types: single-step and multistep procedures. Single-step procedures provide for parallel (simultaneous) testing and simultaneous (adjusted) confidence intervals for assessing the magnitude of the treatment effects. Single-step procedures tend to cause loss of study power, so that sample sizes need to be increased in comparison to sample sizes needed for a single-endpoint study. Multistep procedures are generally more efficient in that they better preserve the power of the tests, but do not readily provide adjusted confidence intervals. There are several kinds of multistep procedures, for example step-down, step-up, and sequential procedures.

In a step-down procedure, one calculates the p-values from all tests to be considered at one time and starts hypothesis testing with the smallest p-value (i.e., statistically the most robust endpoint test) and then steps down to the next smallest p-value (i.e., the next most robust endpoint test), and so on. In a step-up procedure, one proceeds in the reverse direction. That is, one starts with the largest p-value (i.e., the least robust test) and steps up to the second-largest p-value, finally reaching the smallest p-value (i.e., the most robust test). These approaches are covered in the following sections; e.g., the Holm procedure is a step-down procedure and the Hochberg procedure is a step-up procedure.

1. The Bonferroni Method

The Bonferroni method is a single-step procedure that is commonly used, perhaps because of its simplicity and broad applicability. It is a conservative test and a finding that survives a Bonferroni adjustment is a credible trial outcome. The drug is considered to have shown effects for each endpoint that succeeds on this test. The Holm (section IV.C.2) and Hochberg (section IV.C.3) methods are more powerful than the Bonferroni method for primary endpoints and are therefore preferable in many cases. However, for reasons detailed in sections IV.C.2-3, sponsors may still wish to use the Bonferroni method for primary endpoints in order 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 alpha (typically 0.05) equally among the chosen endpoints. The method then concludes that a treatment effect is significant at the alpha 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 alpha for each endpoint is 0.025, with four endpoints it is 0.0125, and so on. Therefore, if a trial with four endpoints produces two-sided p-values of 0.012, 0.026, 0.016, and 0.055 for its four primary endpoints, the Bonferroni method would compare each of these p-values to the divided alpha of 0.0125. The method would conclude that there was a significant treatment effect at level 0.05 for only the first endpoint, because only the first endpoint has a p-value of less than 0.0125 (0.012). If two of the p-values were below 0.0125, then the drug would be considered to have demonstrated effectiveness on both of the specific health effects evaluated by the two endpoints.

The Bonferroni method tends to be conservative for the study overall Type I error rate if the endpoints are positively correlated, especially when there are a large number of positively-correlated endpoints. Consider a case in which all of three endpoints give nominal p-values between 0.025 and 0.05, i.e., all 'significant' at the 0.05 level but none significant under the Bonferroni method. Such an outcome seems intuitively to show effectiveness on all three endpoints, but each would fail the Bonferroni test. When there are more than two endpoints with, for example, correlation of 0.6 to 0.8 between them, the true family-wise Type I error rate may decrease from 0.05 to approximately 0.04 to 0.03, respectively, with negative impact on the Type II error rate. Because it is difficult to know the true correlation structure among different endpoints (not simply the observed correlations within the dataset of the particular study), it is generally not possible to statistically adjust (relax) the Type I error rate for such correlations. When a multiple-arm study design is used (e.g., with several dose-level groups), there are methods that take into account the correlation arising from comparing each treatment group to a common control group.

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, 0.1, 0.3, and 0.2, for four endpoints). These weights are prespecified in the design of the trial, taking into consideration the clinical importance of the endpoints, the likelihood of success, or other factors. There are two ways to perform the weighted Bonferroni test:

- The unequally weighted Bonferroni method is often applied by dividing the overall alpha (e.g., 0.05) into unequal portions, prospectively assigning a specific amount of alpha to each endpoint by multiplying the overall alpha by the assigned weight factor. The sum of the endpoint-specific alphas will always be the overall alpha, and each endpoint's calculated p-value is compared to the assigned endpoint-specific alpha.

- An alternative approach is to adjust the raw calculated p-value for each endpoint by the fractional weight assigned to it (i.e., divide each raw p-value by the endpoint's weight factor), and then compare the adjusted p-values to the overall alpha of 0.05.

These two approaches are equivalent.

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., $\alpha/[m-1]$ for the second smallest p-value, a somewhat less conservative significance level). If $p_{(2)} < \alpha/(m-1)$, 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 $\alpha/(m-2)$, 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 it cannot be concluded that a treatment effect is shown for those remaining endpoints.

For example, when α = 0.05, and there are four endpoints (m = 4), the significance level for the smallest p-value is α/m = 0.05/4 = 0.0125, and significance levels for the subsequent ordered p-values are α/(m-1) = 0.05/3 = 0.0167, α/(m-2) = 0.05/2 = 0.025, and α/(m-3) = 0.05/1 = 0.05, respectively.

To illustrate, we apply the Holm procedure to the two-sided study result p-values used to explain the Bonferroni method: 0.012, 0.026, 0.016, and 0.055 associated with endpoints one to four, respectively (p_1, p_2, p_3, p_4). With four endpoints, the successive endpoint-specific alphas are 0.0125, 0.0167, 0.025, and 0.05. The smallest p-value in this group is p_1= 0.012, which is less than 0.0125. The treatment effect for endpoint one is thus successfully demonstrated and the test continues to the second step. In the second step, the second smallest p-value is p_3 = 0.016, which is compared to 0.0167. Endpoint three has therefore also successfully demonstrated a treatment effect, as 0.016 is less than 0.0167. Testing is now able to proceed to the third step, in which the next ordered p-value of p_2 = 0.026 is compared to 0.025. In this comparison, as 0.026 is greater than 0.025, the test is not statistically significant. This non-significant result stops further tests. Therefore, in this example, this procedure concludes that treatment effects have been shown for endpoints one and three.

As noted, the Holm procedure is less conservative (and thereby more powerful) than the Bonferroni test. It tests the smallest p-value at the same alpha as the Bonferroni test, but, given a statistically-significant result on that endpoint, it tests subsequent p-values at higher significance levels. In the above example, the Bonferroni test was able to conclude that there is a significant treatment effect at the overall level 0.05 for endpoint one only; the Holm test was able to do so for endpoints one and three. Both, however, require at least one endpoint with a p-value < 0.05/*m*. The Holm procedure is also more flexible than simple prospective ordering of endpoints for testing (section IV.C.5). It allows testing of the endpoint with the smallest p-value first, without knowing in advance which endpoint that will be. A disadvantage of the Holm procedure is the potential inability to pass along *unused alpha* (see section IV.C.6) to a secondary endpoint family because testing of any additional endpoints is not permitted when one of the sequentially-tested endpoints in the family fails to reject the null hypothesis.

3. The Hochberg Procedure

The Hochberg procedure is a multi-step, step-up testing procedure. It compares the p-values to the same alpha critical values of α/m, α/(m-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, α/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.

To illustrate, consider the same two-sided p-values used in the previous examples: 0.012, 0.026, 0.016, and 0.055 associated with endpoints one to four, respectively (p_1, p_2, p_3, p_4).

i. The largest p-value of $p_4 = 0.055$ is compared to its alpha critical value of $\alpha = 0.05$. Because this p-value of 0.055 is greater than 0.05, the treatment effect for the endpoint four associated with this p-value is considered not significant. The procedure, however, continues to the second step.

ii. In the second step, the second largest p-value, $p_2 = 0.026$, is compared to $\alpha/2 = 0.025$; p_2 is also greater than the allocated alpha, and endpoint two associated with this p-value is also not statistically significant.

iii. In the third step, the next largest p-value, $p_3 = 0.016$, is compared to its alpha critical value of $\alpha/3 = 0.0167$, and this endpoint does show a significant treatment effect.

iv. The significant result on endpoint three automatically causes the treatment effect for all untested endpoints (which will have smaller p-values) to be significant as well (i.e., endpoint one in this case).

Although for this specific example, the endpoints that are statistically significant are the same as for the Holm procedure, the Hochberg procedure is potentially more powerful. The Hochberg procedure may conclude that there are significant treatment effects for more endpoints than would the Holm procedure, depending on the specific p-values obtained in the study. This is because the Hochberg procedure allows testing of endpoints from the largest p-value to the smallest and concludes that all remaining endpoints are successful as soon as one test is successful, even if the remaining p-values would not have succeeded on testing with their appropriate sequential alpha level. In contrast, the Holm procedure tests from smallest p-value to largest and determines that all untested endpoints are unsuccessful as soon as one test is unsuccessful, even if those remaining endpoints would have been successful if tested with their appropriate sequential alpha level.

Thus, for the case of two endpoints, if the two-sided p-values were 0.026 and 0.045, the Hochberg procedure will conclude that there are significant treatment effects on both endpoints, but the Holm procedure will fail on both. In the Hochberg procedure, the larger of the two p-values, $p = 0.045$ ($< \alpha = 0.05$), is a significant result, and the second endpoint is automatically considered significant. In the Holm procedure, the smaller of the two p-values, 0.026 ($> \alpha/m = 0.05/2$), is a non-significant result; therefore, the larger p-value is not evaluated.

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 and also for two positively-correlated dependent tests with standard test statistics, such as the normal Z, student's t, and 1 degree of freedom chi-square. 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 endpoints. Therefore, beyond the aforementioned cases where the Hochberg procedure is known to be valid, its use is generally not recommended for the primary comparisons of confirmatory clinical trials unless it can be shown that adequate control of Type I error rate is provided.

4. Prospective Alpha Allocation Scheme

The Prospective Alpha Allocation Scheme (PAAS) 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 must receive 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. When the Type I error rate for the study is set at 0.05 overall, the probability of correctly not rejecting any of the individual null hypotheses (i.e., when all null hypotheses are true) must be 1- 0.05 = 0.95 = (1-alpha). 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, must 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 must always satisfy the above equation. In general, when arbitrary alpha allocations are made for some endpoints, at least the last endpoint's alpha must 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 must equal the overall study-wise alpha.

Consider the case of three endpoints with two arbitrary alpha allocations in which $\alpha_1 = 0.02$ and $\alpha_2 = 0.025$ are assigned to the first two endpoints. If the total $\alpha = 0.05$, then the third endpoint would have an alpha of 0.0057, because the above equation becomes $(0.98)(0.975)(1 - \alpha_3) = 0.95$, so that $\alpha_3 = 0.0057$ for the third endpoint, instead of 0.005, as would have been assigned by the Bonferroni method (0.02 + 0.025 + 0.005 = 0.05). When all alpha allocations are equal, then the individual comparison alpha is given by $1 - (1 - \alpha)^{1/m}$. This adjustment formula is also known as the Šidák adjustment formula. For the case of three endpoints, this adjusted alpha is 0.01695, which is only slightly greater than the 0.0167 assigned by the Bonferroni method. The slight savings in alpha provides a slight gain in the power of the tests. The PAAS ensures FWER control for all comparisons that are independent or positively correlated. If the endpoints are negatively correlated, FWER control may not be assured.

5. The Fixed-Sequence Method

The multiplicity problem arises from conducting tests for each of the multiple endpoints where each test provides an opportunity to decide that the study was successful. Any method that adequately adjusts for the multiplicity of opportunities will address the problem. 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 strategy tests endpoints in a predefined order, all at the same significance level alpha (e.g., $\alpha = 0.05$), moving to a second endpoint only after a success on the previous endpoint. Such a test procedure does not inflate the Type I error rate as long as there is (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 idea behind this sequential testing method is that when there is a significant treatment effect for an endpoint, then the alpha level for this test remains available to be carried forward (passed along) to the next endpoint test in the sequence. However, the method uses all of the available alpha as soon as a non-significant result occurs. The order of testing is therefore critical.

The statistical conclusions provided by this method may differ from those provided by other methods, and they depend on the ordering of the tests. Consider, for example, a trial with three primary endpoints, A, B, and C, whose two-sided *p*-values for treatment effects are: $p_A = 0.045$, $p_B = 0.016$ and $p_C = 0.065$. This trial would conclude that there was a significant treatment effect for only the endpoint B by the Bonferroni test, because $p_B = 0.016 < 0.0167$ (i.e., 0.05/3), but would not conclude that there was a significant effect on endpoints A or C. The Holm test would not find significant effects for additional endpoints either, unless the p-value for endpoint A was $p < 0.025$. If the study had planned sequential testing in the order of (C, B, A), it would be an entirely failed study, because $p_C = 0.065 > 0.05$, and no further testing would be performed after the first failed test for endpoint C. On the other hand, this trial would show significant treatment effects for endpoints B and A if it had planned sequential testing in the order of (B, A, C), because $p_B = 0.016 < 0.05$, and following it, $p_A = 0.045 < 0.05$; the same effects would be shown if the order was (A, B, C). Thus, the fixed-order sequential testing method has the potential to find more endpoints successful than the single-step methods, but it also has the potential to find fewer endpoints successful, depending on the order chosen.

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, a statistical conclusion cannot be made about 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 $p = 0.250$, and the p-value for the second endpoint is $p = 0.0001$; despite the apparent "strong" finding for the second endpoint, no formal favorable statistical conclusion can be reached for this endpoint. Although it may seem counterintuitive to ignore such an apparently strong result, to allow a conclusion of drug effectiveness based on the second endpoint would in fact be ignoring the first endpoint's result and returning to the situation of having multiple separate opportunities to declare the study a success. Such a post hoc rescue recreates the multiplicity problem, and causes inflation of the study-wise Type I error rate. The example discussed here would, of course, have shown an effect using a Bonferroni test.

Thus, carefully selecting the ordering of the tests of hypotheses is essential. 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, and there are other methods for addressing the multiplicity problem, which are described in the following subsections.

6. The Fallback Method

The fallback method is a modification of the fixed-sequence method that provides some opportunity to test an endpoint later in the sequence even if an endpoint tested early in the sequence has failed to show statistical significance. The order of the endpoints remains important. The appeal of the fallback method is that if an endpoint later in the sequence has a robust treatment effect while the preceding endpoint is unsuccessful, there is a modest amount of alpha retained as a fallback to allow interpretation of that endpoint without inflating the Type I error rate.

Applying the fallback method begins by dividing the total alpha (not necessarily equally) among the endpoints, and maintains a fixed sequence for the testing. As the testing sequence progresses, a successful test preserves its assigned alpha as "saved" (unused) alpha that is passed along to the next test in the sequence, as is the case for the sequential method. This passed-along alpha is added to the prospectively assigned alpha (if any) of that next endpoint and the summed alpha is used for testing that endpoint. Thus, as sequential tests are successful, the alpha accumulates for the endpoints later in the sequence; these endpoints are then tested with progressively larger alphas.

To illustrate, consider a cardiovascular trial in which the first primary endpoint is exercise capacity, for which the trial is adequately powered. The second primary endpoint is mortality, for which the trial is underpowered.

i. Under the fallback method, we may assign $\alpha_1 = 0.04$ for the first endpoint test and save alpha of 0.01 for the second endpoint test. Any other desired division of the available overall alpha would also be permitted.

ii. If the first endpoint test is significant at level $\alpha_1 = 0.04$, this alpha is unused and is passed to the second endpoint test as an additional alpha of 0.04, giving a total alpha for the second endpoint test of 0.05 (0.01 + 0.04). The second endpoint test is then performed at the significance level of 0.05.

iii. If the first endpoint is not significant at level 0.04, then this alpha of 0.04 is not available to be passed on for the second endpoint test. The test for the second endpoint is at the originally reserved alpha of 0.01.

In practice, users of this method usually assign most of the alpha to the first primary endpoint and the remainder to the second endpoint, although other distributions are also valid. The fallback method is often used when there is an endpoint thought less likely than another to be statistically significant, so that it is not designated the first endpoint, but is nevertheless of substantial clinical importance. The fallback method could conclude that an unexpectedly robust finding is statistically interpretable as a positive result even if the first primary endpoint failed, without inflation of the Type I error rate.

The statistical power of the fallback method depends primarily on the magnitude of the effect on, and alpha assigned to, each of the ordered endpoints. As with the simple fixed-sequence method, the overall power of the fallback method exceeds that of the Bonferroni test, because when the earlier endpoints show significant results, the method uses larger alpha levels for later endpoints than is possible under the Bonferroni method.

7. Gatekeeping Testing Strategies

Clinical trials commonly assess efficacy of a treatment on multiple endpoints, usually grouped into a primary endpoint or endpoint family, and a secondary endpoint or endpoint family (see sections II.A and III.A). The usual strategy is to test all endpoints in the primary family according to one of the previously discussed methods (e.g., Bonferroni, fallback) and proceed to the secondary family of endpoints only if there has been statistical success in the primary family. This allows all of the available alpha level to be distributed within the primary family (containing the most important study endpoints) and thus maximizes the study power for those endpoints. In contrast, if the available alpha were distributed among all of the endpoints in the primary and secondary families, power would be reduced for the primary endpoints. Although it is not generally recommended, if there were an additional family of endpoints for which it was also important to control the Type I error rate, that family could be designated as third in the sequence.

This approach of testing the primary family first, and then the secondary family contingent upon the results within the primary family is called the gatekeeping testing strategy to highlight the fact that the endpoint families are analyzed in a sequence, with each family serving as a gatekeeper for the next one. The tests for the secondary family (and subsequent families if any) are carried out with appropriate multiplicity adjustments within that family, but only if the tests in the primary family have been successful.

Two types of gatekeeping testing strategies are common in clinical trials, serial and parallel, determined by how the endpoints are tested within the primary family. The term serial strategy is applied when 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 same alpha level (e.g., $\alpha = 0.05$), the endpoints in the second family are examined. The endpoints in the second family are tested by any one of several possible methods (e.g., Holm procedure, the fixed-sequence method, or others described in section IV.C). 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.

The term parallel gatekeeping strategy is applied when the endpoints in the primary family are not all co-primary endpoints, and a testing method that allows the passing along of alpha from an individual test to a subsequent test (e.g., Bonferroni method or Truncated Holm method described next) is specified. In this strategy, the second endpoint family is examined when at least one of the endpoints in the first family has shown statistical significance.

The Bonferroni method is sometimes used for the parallel gatekeeping strategy, as it is the simplest approach. The secondary endpoint family may use a different method (e.g., the fixed-sequence method or Holm method). In this approach, if an endpoint comparison within the primary family is statistically significant at its allocated (or accumulated) endpoint-specific alpha level, then this alpha level can be validly passed on to the next family. On the other hand, if an endpoint comparison in a family is not significant at its endpoint-specific alpha level, that alpha is not passed on to the next family. The overall alpha available for testing the secondary family is the accumulated (unused) endpoint-specific alpha levels of those comparisons in the primary family that were found significant.

To illustrate, consider a trial whose primary objective is to test for superiority of a treatment to placebo for five endpoints: A, B, C, D and E. For this objective, the trial organizes the endpoints hierarchically into a primary family F1 = {A, B} and a secondary family F2 = {C, D, and E}. The statistical plan is to assign the total available alpha (0.05) to F1 and test the endpoints A and B in F1 by the Bonferroni method at endpoint-specific alpha levels of 0.04 and 0.01, respectively. No alpha is reserved for the second family, and the second family is tested with the Holm procedure with whatever amount of alpha is passed along to it. If, at the completion of the tests for F1, the p-values for the endpoints A and B are 0.035 and 0.055, respectively, and the p-values for endpoints C, D and E are 0.011, 0.045, and 0.019, respectively, then:

i. The result for endpoint A is significant, but the result for endpoint B is not, leaving alpha of 0.04 as unused and alpha of 0.01 as used.

ii. The total alpha available for testing the endpoints in F2 is 0.04 and not 0.05.

iii. The endpoints C and E are significant at level 0.04 by the Holm test (C, E, and D are tested at levels of 0.0133, 0.02, 0.04, respectively).

The gatekeeping method described above controls the study-wise Type I error rate (e.g., at level 0.05) associated with the trial's primary and secondary families. The study-wise Type I error rate takes into consideration the potential for an erroneous conclusion of efficacy for any endpoint in any family and the multiple possibilities of the drug being truly effective or ineffective on any of the endpoints. The gatekeeping strategy controls the study-wise Type I error rate when the principle of passing along only unused alpha from statistically-significant tests of hypotheses is applied. In contrast, however, independent error rate control of each family's FWER (i.e., testing each family at a separate 0.05) can lead to inflation of the study-wise Type I error rate when some, but not all, of the null hypotheses for the primary endpoint family are in fact true.

8. The Truncated Holm and Hochberg Procedures for Parallel Gatekeeping

When used as a gatekeeping strategy to test the primary family of endpoints, the Bonferroni method and some other single-step methods (such as the Dunnett's test, which is not covered in this document) have an important property of preserving some alpha for testing the secondary endpoint family when at least one of the endpoints in the primary family is statistically significant. In the Bonferroni method, the endpoint-specific alpha from each test that successfully rejected that null hypothesis is summed and becomes the alpha available to the secondary endpoint family. For example, in the equally weighted Bonferroni method, when there are two endpoints in the primary family, the unused alpha available for tests of hypotheses in the secondary family can be 0.05, 0.025, or 0, depending, respectively, on whether both, one, or none of the primary endpoint tests rejected their respective null hypotheses.

The conventional Holm and Hochberg methods, however (see sections IV.C.2 and IV.C.3), do not have this property. These methods pass alpha from the primary family to the secondary family only when all of the null hypotheses in the primary family are rejected. These two methods give better power on recycling all alpha within the family and releasing it only when all hypotheses in that family are rejected. Inappropriately proceeding as if there is some preserved alpha when a study fails to reject one or more of the primary hypotheses will result in an inflated overall Type I error rate.

There are, however, procedures called the truncated Holm and the truncated Hochberg that can be used when there is a desire to have the power advantage of the conventional Holm or Hochberg procedures but also to have some alpha available for testing the secondary endpoint family if at least one of the primary endpoints is successful. In a truncated Holm or Hochberg procedure, some portion of the unused alpha from each step is reserved for passing to the secondary endpoint family. The truncated Holm procedure and the truncated Hochberg procedures are hybrids of their conventional forms and the Bonferroni method. As a consequence, the endpoint-specific alpha for each successive test of hypothesis of the primary endpoints after the first is not as large as in the conventional Holm or the conventional Hochberg procedure. In either of these approaches, of course, if all of the individual endpoint tests of hypotheses in the primary endpoint family successfully reject the null hypothesis, the full alpha of 0.05 is available for the secondary endpoint family. The amount of reserved alpha from the successive tests should be chosen carefully, as the choice creates a balance between decreasing study power for the endpoints in the primary family and the guarantee (if at least the first test rejects the null hypothesis) of some power to test the secondary endpoint family. The following example illustrates these two procedures for a primary family with three endpoints.

Consider treatment versus control comparisons for three endpoints in the primary family with the control of alpha at the 0.05 level. The endpoint-specific alpha levels for the conventional Holm for this case are 0.05/3, 0.05/2, and 0.05 (see section IV.C.2), and those by the equally weighted Bonferroni method are 0.05/3, the same for each comparison (see section IV.C.1). The endpoint-specific alpha levels for the truncated Holm are then constructed by combining the endpoint-specific alpha levels of the two methods with a “truncation fraction” of *f*, whose value between zero and one is selected in advance. The following calculations illustrate this combination using *f* =1/2; the multipliers with *f* are the endpoint-specific alpha levels for the conventional Holm and those with (1-*f*) are by the equally weighted Bonferroni method.

$$\alpha_1 = \frac{0.05}{3} f + \frac{0.05}{3}(1-f) = \frac{0.05}{3}.\frac{1}{2} + \frac{0.05}{3}(1-\frac{1}{2}) = 0.0167$$

$$\alpha_2 = \frac{\frac{0.05}{2} f + \frac{0.05}{3}(1-f)}{} = \frac{0.05}{2}.\frac{1}{2} + \frac{0.05}{3}(1-\frac{1}{2}) = 0.0208$$

$$\alpha_3 = \frac{0.05}{1} f + \frac{0.05}{3}(1-f) = \frac{0.05}{1}.\frac{1}{2} + \frac{0.05}{3}(1-\frac{1}{2}) = 0.0333$$

Thus, for this particular case, when the value of $f = 1/2$, the first test for the truncated Holm test is performed at $\alpha_1 = 0.0167$, which is the same for the conventional Holm test. However, the second test, after the first test is successful, is performed at level $\alpha_2 = 0.0208$, and the third test, after the first two tests are successful, is at level $\alpha_3 = 0.0333$. The unused alpha levels for passing to the secondary family are calculated as:

i. Unused alpha = 0.05, if all three tests are successful;
ii. Unused alpha = $(0.05 - \alpha_3) = 0.05 - 0.0333 = 0.0167$, if the first two tests are successful, but the last one is not;
iii. Unused alpha = $(0.05 - 2\alpha_2) = 0.05 - 2(0.0208) = 0.0084$, if the first test is successful, but the other two tests are not.

For the truncated Hochberg, alpha levels α_1, α_2, and α_3 are the same as those for the truncated Holm, except that for the truncated Hochberg, the first test starts with the largest p-value (i.e., largest of the three endpoint treatment-to-control comparison p-values) at level $\alpha_3 = 0.0333$. If this first test is successful, then the other two tests are also considered successful, and alpha of 0.05 passes to the secondary family. However, if the first test is not successful, then the second test with second-largest p-value is at level $\alpha_2 = 0.0208$. If this second test is successful, then the remaining last test is also considered successful, and alpha of 0.0167 passes to the secondary family. However, if this second test is not successful, then the last test with the smallest p-value is at level $\alpha_1 = 0.0167$, and if that test is successful, then alpha of 0.0084 passes to the secondary family. This illustration is with $f = 1/2$. Similar calculations would follow for different values of f.

9. *Multi-Branched Gatekeeping Procedures*

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). Each individual hypothesis to test pertains to one particular endpoint, dose, and analysis objective. The total number of hypotheses is the product of the number of options within each dimension and can become large, even when there are only two or three options for each dimension.

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. After demonstrating non-inferiority on the endpoint, it is possible to then test for superiority at an unadjusted alpha. In a fixed-sequence (unbranched) approach, it would also be appropriate to analyze a second endpoint for non-inferiority at the same alpha after the first endpoint is successfully shown to be non-inferior.

Suppose, however, that one wants to carry out both of these analyses after showing non-inferiority for the first endpoint. 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. There is a choice of statistical adjustments to apply in this setting.

Treating the hypotheses as independent and applying a simple method such as Bonferroni leads to testing these hypotheses at small alpha levels, and consequently a very large study may be necessary to ensure good study power. Alternatively, applying a fixed-sequence method may lead to many endpoint tests being disallowed because the optimal sequence for testing is usually not prospectively determinable. The multi-branched gatekeeping procedure can address multiplicity problems of this multi-dimensional type. 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 1), consider a clinical trial that compares a treatment to control on two primary endpoints (endpoint one and endpoint two) 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.

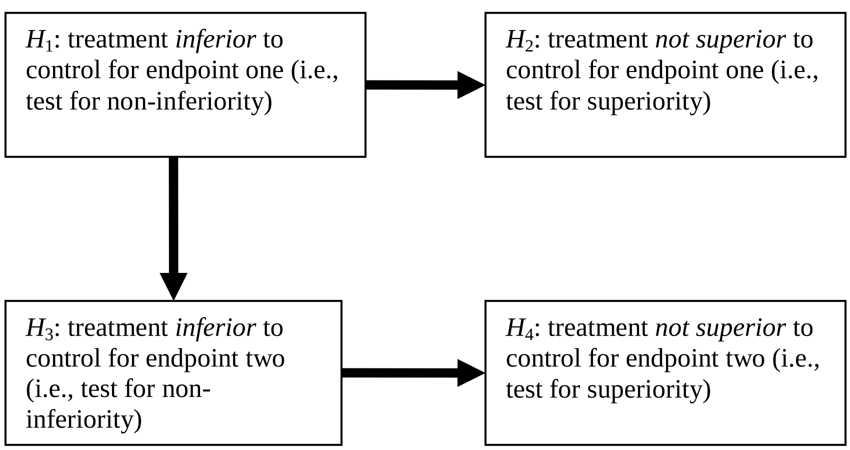


Figure 1: Example of a flow diagram for non-inferiority and superiority tests for endpoints one and two of a trial with logical restrictions: in order to test for superiority for endpoint one and/or two, one must first establish non-inferiority for that endpoint.

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. A different study might have three dimensions, two endpoints to be tested at two dose levels (along with a control group) with non-inferiority and superiority tests on each endpoint, having restrictions, e.g., that the lower dose can be tested after a success on the higher dose, and superiority on an endpoint can be tested after non-inferiority has been shown.

For the test strategy in Figure 1, one may, inappropriately, test each hypothesis at the same significance level (e.g., $\alpha = 0.05$), reasoning that the tests for non-inferiority for the two endpoints follow a sequential order, allowing passing along the full alpha; and that the test for superiority for each endpoint follows naturally after non-inferiority for it is first demonstrated. This approach, however, is likely to inflate the overall Type I error rate, because in Figure 1, the testing path (sequence) after the node at H_1 splits into two branches; one goes on to test for H_2 and the other to test for H_3. Consequently, once the trial concludes non-inferiority of the treatment to control for endpoint one, erroneous conclusions for tests of H_2 and H_3 can occur in multiple ways; that is, either H_2 is erroneously rejected, or H_3 is erroneously rejected, or both H_2 and H_3 are erroneously rejected. If each of these separate hypotheses were to be tested at the 0.05 level, this would obviously lead to Type I error rate inflation. As another illustration of Type I error rate inflation, suppose that in reality the treatment is non-inferior to control for both endpoints but is not superior to control for either endpoint. In this scenario, the testing scheme (without alpha adjustments) can conclude superiority of the treatment to control in multiple ways, i.e., the treatment is superior to control for either endpoint one or endpoint two, or for both endpoints.

It is possible to deal with this problem using the Bonferroni-based gatekeeping method by grouping the hypotheses as follows:

- Group one includes only H_1 (the test of non-inferiority for endpoint one)
- Group two includes H_2 (the test of superiority for endpoint one) and H_3 (the test of non-inferiority for endpoint two)
- Group three includes only H_4 (the test of superiority for endpoint two).

The procedure would begin with the test of the single hypothesis H_1 in group one at the level intended for the study-wise overall Type I error rate (e.g., $\alpha = 0.05$). Group one serves as a gatekeeper for group two. Therefore, once the result for H_1 is significant at level α (i.e., the treatment is non-inferior to control for endpoint one at level α), testing proceeds to the hypotheses H_2 and H_3 in group two with the alpha that was not used within family one, which in this case would be the overall study alpha.

The test of H_2 and H_3 in family two can use the Bonferroni method at the endpoint-specific alpha of 0.025 for each test according to the Bonferroni-based gatekeeping method. The standard Holm procedure is not considered here for the reason discussed in sections IV.C.2 and IV.C.8. Dividing the available alpha between the two endpoints will reduce study power for these endpoints (or necessitate an increased sample size to maintain study power), making it more difficult for the study to succeed on these endpoints; but it is necessary to maintain control of Type I error rate.

Therefore, if both H_2 and H_3 are rejected, H_4 is tested at $\alpha = 0.05$. However, if only H_3 is rejected, then H_4 is tested at $\alpha = 0.025$. If H_3 is not rejected but H_2 is rejected, H_4 could be tested at $\alpha = 0.025$ in accord with the plan, but this would be illogical because if endpoint two failed to show non-inferiority (H_3), superiority could not have occurred.

When there are three or more dimensions and multiple branch points, planning the sequence of testing becomes complex and difficult to describe in the manner illustrated here. In these situations, the graphical approach to displaying and evaluating analysis paths (Appendix A) can be valuable.

10. Resampling-Based, Multiple-Testing Procedures

When there is correlation among multiple endpoints, resampling 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.

There is at present little experience with these methods in drug development clinical trials. Because of this, resampling methods are not recommended as primary analysis methods for adequate and well-controlled trials in drug development. It may, however, be useful and instructive to compare the results of resampling methods with those obtained using conventional methods in order to gain experience with and understanding of resampling methods' properties, advantages, and limitations.

V. CONCLUSION

The chance of making a false positive conclusion, concluding that a drug has a beneficial effect when it does not, is of primary concern to FDA. The widely accepted standard is to control the chance of coming to a false positive conclusion (Type I error probability) about a drug's effects to less than 2.5 percent (1 in 40 chance). As the number of endpoints or analyses increases, the probability of making a false positive conclusion can increase well beyond the 2.5 percent standard. Multiplicity adjustments, as described in this guidance, provide a means for controlling Type I error when there are multiple analyses of the drug's effects. There are many strategies and/or choices of methods that may 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 to be addressed at the study-planning stage. Statistical expertise should be enlisted to help choose the most appropriate approach. Failure to appropriately control the Type I error rate can lead to false positive conclusions; this guidance is intended to clarify when and how multiplicity due to multiple endpoints should be managed to avoid reaching such false conclusions.

GENERAL REFERENCES

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

Bauer P. Multiple testing in clinical trials. *Statistics in Medicine* 1991; **10**: 871-890.

Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*, CRC Press (Taylor & Francis Group), Chapman and Hall, 2010.

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

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

Chi GYH. Some issues with composite endpoints in clinical trials. *Fundamental & Clinical Pharmacology* 2005; **19**: 609-619.

CPMP/EWP/908/99. Points to consider on multiplicity issues in clinical trials. September 2002; http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf.

Dmitrienko A, Tamhane AC, Bretz F. *Multiple testing problems in pharmaceutical statistics*, CRC Press (Taylor & Francis Group), Chapman & Hall/CRC Biostatistics Series, 2010.

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, Tamhane AC. *Multiple Comparison Procedures*. John Wiley & Sons, New York, 1987.

Holm SA. A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics* 1979; **6:** 65-70.

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

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

Huque MF. Validity of the Hochberg procedure revisited for clinical trial applications. *Statistics in Medicine* 2015, (wileyonlinelibrary.com) DOI: 10.1002/sim.6617.

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

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

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

Moye LA. *Multiple Analyses in Clinical Trials*. Springer-Verlag, New York, 2003.

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

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

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

Westfall PH, Tobias RD, Rom D, Wolfinger RD, HochbergY. *Multiple Comparisons and Multiple Tests Using the SAS® System*, SAS Institute Inc.: Cary, NC, USA, 1999.

Westfall PH, Young SS. *Resampling Based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons, Inc. New York, 1993.

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

REFERENCES TO EXAMPLES

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

Dahlöf G, Devereux RB, Kjeldsen SE, Julius S, Beevers G, de Faire U, Fyhrquist F, Ibsen H, Kristiansson K, Lederballe-Pedersen O, Lindholm LH, Nieminen MS, Omvik P, Oparil S, Wedel

H: LIFE Study Group. Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. *Lancet* 2002; 359(9311): 995-1003.

APPENDIX: THE GRAPHICAL APPROACH

A graphical approach is available for developing and evaluating hierarchal multiple analysis strategies. This approach provides a means for specifying, communicating, and assessing different hypothesis testing strategies, but is not by itself an additional method for addressing multiplicity (such as those described in section IV). Instead, the graphical approach is a means of depicting a strategy consisting of the previously described Bonferroni-based sequential methods, such as fixed-sequence, fallback type, and gatekeeping procedures. 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. This technique will be most helpful when the analysis plan is complex due to splitting of the overall alpha among several endpoints (either initially or after a particular endpoint has been successful), particularly if there is a desire to have a second chance for an endpoint that was not statistically significant at the initially assigned endpoint-specific alpha, but can receive pass-along alpha from a different endpoint that was successful (the loop-back feature described below). This situation may occur when complex testing strategies are being considered because of intricate endpoint relationships and differing endpoint importance.

Graphical displays of complex analysis strategies can aid in clearly describing and assessing the proposed plan by displaying all the logical relationships among endpoint tests of hypotheses. In addition, simple modifications of the initial graph can easily create different variations of a test strategy, aiding comparison among the variations. The graphical approach can be useful in trial design to identify a test scheme that is suitably tailored to the objectives of the trial.

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 that shows each of the hypotheses (H_1, H_2,..., H_m) located at a vertex (or node, a junction of testing order paths), and depicts the test order paths by lines (with the direction of the path indicated by an arrowhead) connecting the hypotheses. Each vertex (hypothesis) is allocated an initial amount of alpha, which we call here 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 study-wise Type I error rate). An exception can occur if one designates two or more hypotheses as a co-endpoint group, so that the same endpoint-specific alpha is applied to all tests in that group.

Each test order path is also assigned a value between 0 and 1, called a weight for that path 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 must 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. With this technique there is no need to explicitly designate hypotheses as part of the primary or secondary endpoint families; more nuanced hierarchies are able to be achieved based on the initial allocation of the endpoint-specific alpha and the division of passed-forward alpha among the test paths leaving each vertex. Clearly, the hypotheses that receive an initial endpoint-specific alpha allocation of 0.0 will often be those regarded as of lesser importance, which is implicitly similar to designating the associated endpoint as a secondary endpoint.

Adhering to the principles outlined in prior sections of this guidance, when an endpoint test is successful in rejecting the corresponding null hypothesis, that endpoint-specific alpha can be passed on to the next test indicated by the arrow, and will be divided among several subsequent hypotheses when there are several paths leaving that vertex. This shift of alpha occurs only when the test result for the hypothesis associated with a vertex at the arrow's tail is significant. Thus, as with the simple fallback method, the actual endpoint-specific alpha used in an endpoint test cannot be determined until the study results are complete and hypothesis testing begins; the sequential test determines which vertices are associated with alpha levels that can be passed along for accumulation in the subsequent test and which are not.

Several examples of the graphical method follow to help illustrate the concept, construction, interpretation, and application of these diagrams. The first several of these examples are simple cases where the graphical approach is no more useful than a nondiagramatic (written text) description, but where the principles of the approach can be more clearly illustrated.

Fixed-Sequence Method

The fixed-sequence testing strategy (section IV.C.5), shown in Figure A1, 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. Arrows indicate 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.

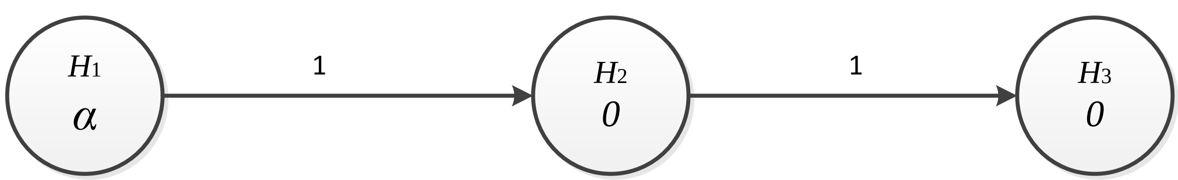


Figure A1: Graphical illustration of the fixed-sequence testing with three hypotheses.

Loop-Back Feature to Indicate Two-Way Potential for Alpha Passing

Another valuable feature of the graphical method occurs when the available alpha level is split between two or more endpoints into endpoint-specific alphas levels; these diagrams can easily illustrate the potential for loop-back passing of endpoint-specific alpha. If a hypothesis is not rejected at its endpoint-specific alpha level, but a different hypothesis is, then the unused endpoint-specific alpha from the rejected second hypothesis can be directed to loop back to the first hypothesis, which is then re-tested at the higher alpha level. Thus, in Figure A2, if assigned endpoint-specific alpha levels for testing H_1 and H_2 are $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$, respectively, and if H_1 is not rejected but H_2 is rejected, then the unused alpha of 0.01 for H_2 loops back to H_1 for re-testing at the higher level of $0.04 + 0.01 = 0.05$. Without the loop-back from H_2 to H_1, this would simply be the fallback method (described in section IV.C.6).

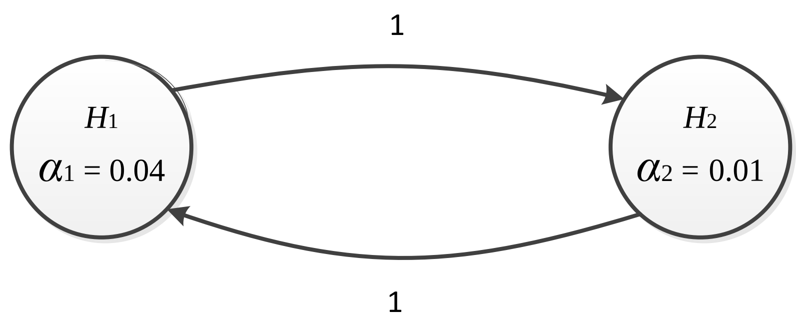


Figure A2: Graphical illustration of the loop back passing of endpoint-specific alpha from H_2 to H_1.

The Holm procedure (section IV.C.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.

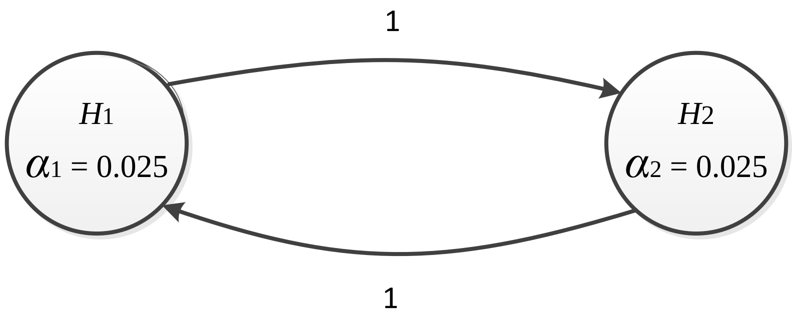


Figure A3: Graphical illustration of the Holm procedure with two hypotheses.

Because of the loop-back procedure and potential for retesting at a larger accumulated endpoint-specific alpha, the figure shows that there is no need for the Holm procedure's rule of starting with the smaller p-value. Testing can begin at either vertex because the other vertex can always be tested, and the first vertex can be retested if it did not succeed on first examination. Both will have an endpoint-specific alpha of at least 0.025, and if one vertex's test is successful, the other hypothesis will be tested (or retested) at the full alpha of 0.05. This is a general principle for analysis strategies described with the graphical approach. Testing on the diagram with loop-back may 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). Testing then follows the arrows, passing the alpha along as directed in the diagram. The final conclusions of which hypotheses were statistically significant 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.

An Improved Fallback Method

Figure A4 (a) displays the conventional fallback test (section IV.C.6) with three hypotheses. Each of the hypotheses is assigned an endpoint-specific alpha so that their sum $\alpha_1 + \alpha_2 + \alpha_3 = \alpha$. If the test result for H_1 is significant, then its level α_1 is passed on to H_2, as indicated by the arrow going from H_1 to H_2. Furthermore, if the test result for H_2 is now significant at its endpoint-specific alpha level (which will be either α_2 or $\alpha_1 + \alpha_2$), then this level is forwarded to H_3 as indicated by the arrow going from H_2 to H_3. Thus, if test results for both H_1 and H_2 are significant, then the total alpha level available for the test of H_3 is $\alpha_1 + \alpha_2 + \alpha_3 = \alpha$.

Examination of the conventional fallback method suggests an improvement, as shown in Figure A4 (b). In the conventional scheme, if the test result for H_3 is significant, then its endpoint-specific alpha level is not shifted to any other hypothesis. Hypothesis H_3, however, is permitted to be tested even if the test of H_2 were not successful. In the case where the test result for H_3 is significant, its endpoint-specific alpha level can be re-used either by H_1 or H_2 or both (if loop-back of the endpoint-specific alpha level of H_3 was divided between H_1 and H_2). Thus, two loop-back arrows can be added to the conventional fallback figure to show the potential for passing back of some portion of H_3's endpoint-specific alpha to H_1, H_2, or both. The actual fraction to be passed back to H_1, and the fraction to H_2, should be prospectively specified, and cannot be adjusted after the study results are examined (when it could be seen which of the two earlier endpoints might most benefit from this passing-back of alpha). Figure A4(b) shows this procedure with the fraction r of this un-used alpha shifted to H_2 and the remaining fraction 1- r of this alpha shifted to H_1. The value of r should be prospectively specified in the study analysis plan.

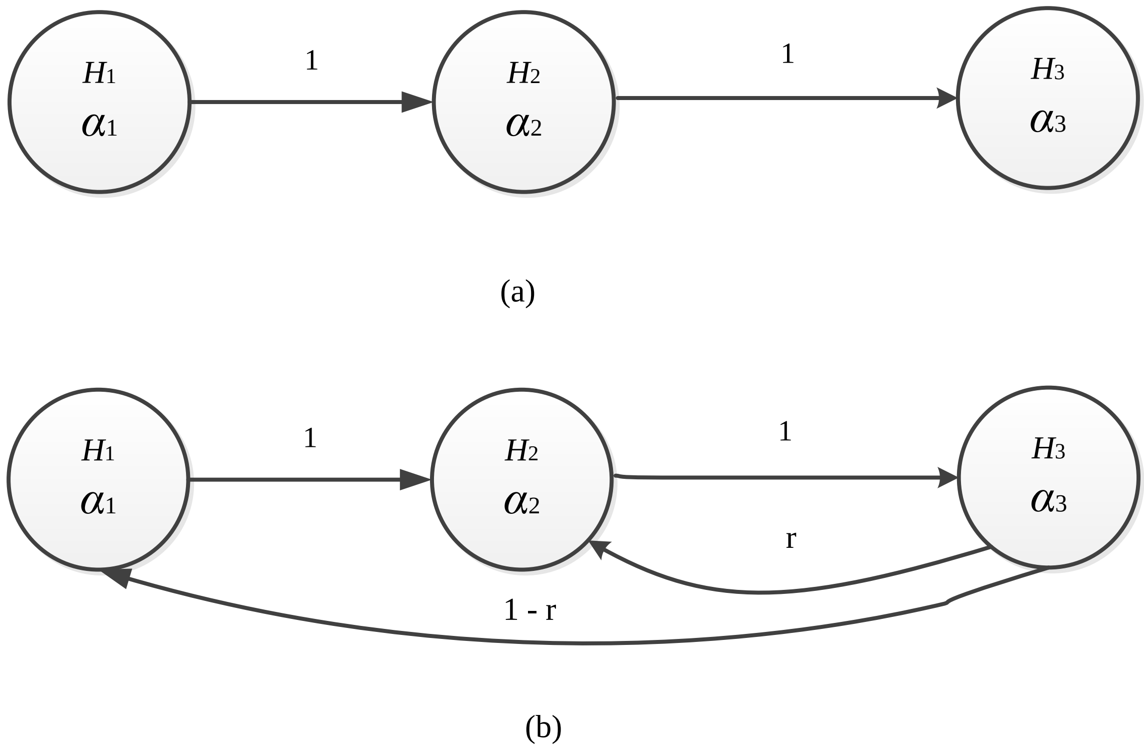


Figure A4: Fallback (a) and improved fallback (b) procedures.

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.

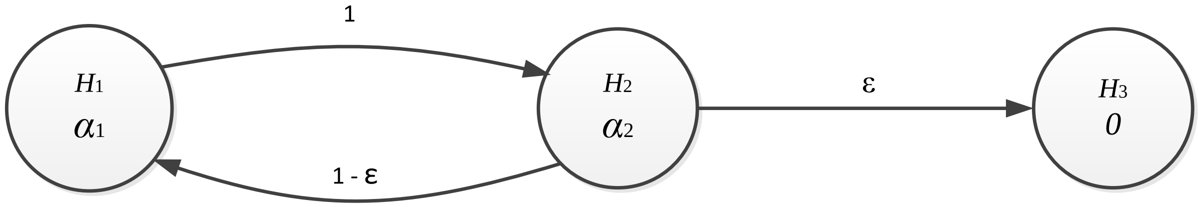
When there is a desire to consider analysis strategies with complex division of alpha, the graphical method and progressive updating of the diagram can aid in understanding the implication of the different strategies for a variety of different hypothetical scenarios. This progressive updating can aid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

Figure A5 is an example of how the graphical method aids in formulating the testing of three hypotheses H_1, H_2, and H_3 and illustrates the updating of the diagram when a test of hypothesis is successful. For this example, the analysis plan designated two hypotheses, H_1 and H_2, to be of prime importance (i.e., primary endpoints), and H_3 (the secondary endpoint) is tested only if the test results for H_1 and H_2 are both significant. Assume that it is desired to always be able to test both H_1 and H_2 (i.e., a willingness to split the available alpha between them), but that if either H_1 or H_2 is successfully rejected, the alpha level of that test would be passed to the other hypothesis if needed, so that it can be tested at the maximal possible alpha level (i.e., the fallback method is specified for the two important endpoints, with α_1 assigned to begin testing on H_1, and α_2 reserved as the minimum that will be available for testing H_2). Thus, as shown in Figure A5 (a), if the test result for H_1 is significant, then its endpoint-specific alpha level α_1 is passed to H_2, so that H_2 is tested at an endpoint-specific alpha level of $\alpha_1 + \alpha_2 = \alpha$. On the other hand, if the test result for H_1 is not significant, H_2 is still tested with the reserved α_2. In this case, however, if the test result for H_2 is significant, the alpha level of α_2 is recycled back to re-testing of H_1 at level $\alpha_1 + \alpha_2 = \alpha$. Note that the graphical method aids in communicating that the re-testing of H_1 at an increased endpoint-specific alpha is part of the prospective analytic plan.

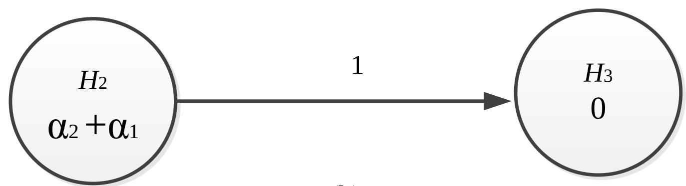
The intended analysis, however, is that if, and only if, these tests of hypotheses (including potential re-test with passed alpha) have successfully rejected H_1 and H_2, then the full available alpha would be passed to H_3. This conditional passing of alpha is depicted by a path from H_2 to H_3 with weight ε. At the start (before any testing of any hypothesis) ε is set to a negligible amount. Because of this, even though there is a path from H_2 to H_3, when H_1 has not yet been successfully rejected, essentially all of α_2 will be passed back to H_1 as the priority over H_3. This scheme will eventually allow for meaningful testing of H_3 if appropriate according to the sequentially updated diagrams.

Figure A5 (b) shows the updated graph when the result for H_1 in Figure A5 (a) is significant at level α_1 and prior to testing H_2 at the now accumulated endpoint-specific alpha of $\alpha_1 + \alpha_2$ (which would be equal to the total alpha for the study in this case). Note that the weight on the path from H_2 to H_3 is now set to 1. This occurs because diagram updating is done when a test of hypothesis is significant. The process of diagram updating first passes along the retained alpha from the successful hypothesis (vertex) according to the weights on the arrows leaving that vertex. That vertex is then eliminated from the diagram and a new diagram is constructed by connecting all the incoming paths (arrows) to all outgoing paths (tails) of the now deleted vertex, and adjusting the pathway weights. The new weights on the new paths are determined based on the relative weights of each previous part of the new path. The essential principle of readjustment of the pathway weights is that the sum of the weights on the outgoing paths from each vertex must be 1.0. This rule causes the weight on the path from H_2 to H_3 to become 1 (from the prior negligible fraction ε) because it is the only remaining path leaving H_2. In some strategies, a newly created connection path arising from elimination of a successful vertex will duplicate a preexisting direct connection between two vertices; in this case the weights of the duplicate paths are combined and drawn as a single path.

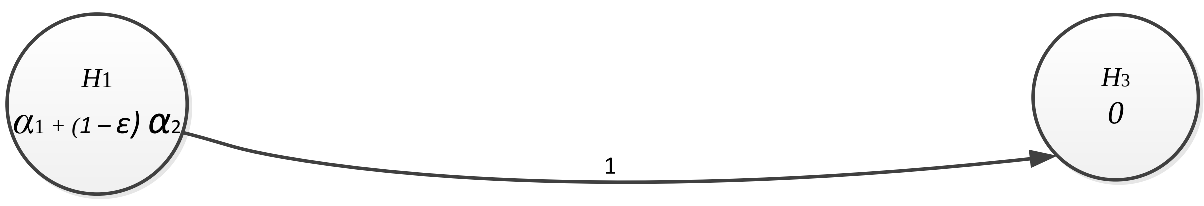
Continuing with the example depicted in Figure A5, if H_1 is not initially significant and H_2 is significant at level α_2, Figure A5 (c) shows the updated diagram prior to re-testing H_1 at the now accumulated endpoint-specific alpha. The vertex for H_2 was eliminated from the updated diagram, and the direct path from H_1 to H_3 is displayed. Both Figures A5 (b) and A5 (c) indicate that H_3 can be tested at the full level α (= $\alpha_1 + \alpha_2 + 0$) when the test results for H_1 and H_2 are both significant, but that no alpha is passed to H_3 unless both H_1 and H_2 were significant.



(a)



(b)



(c)

Figure A5: Graphical illustration of the fallback procedure applied to three hypotheses when the first two hypotheses are most important and the third hypothesis is tested only when both of the first two hypotheses are significant.

(a) The initial diagram shows all hypotheses and paths. The notation ε indicates a positive number close to zero. This convention indicates the potential to pass alpha to H_3, but only if it is not necessary to pass alpha from H_2 to H_1 (see text for explanation).

(b) The updated diagram shows the case where only H_1 was tested and shown to be statistically significant.

(c) The updated diagram shows the case where H_2 was the first hypothesis to be statistically significant at the initially allocated endpoint-specific alpha.

A detailed algorithm for iteratively updating the graph when a test is found significant is illustrated with the final example. Updating of a graph involves determining new endpoint-specific alpha levels and path weights based on satisfying the conditions that (1) the sum of all endpoint-specific alpha levels equals α and (2) the sum of all weights on outgoing arrows from a vertex to others equals 1.0.

The case of three hypotheses with fixed weights on the paths between the hypotheses will be used to illustrate the algorithm (Figure A6 (a)). Suppose that hypothesis H_3 is rejected. The graph needs to be updated to remove this hypothesis and retain hypotheses H_1 and H_2. Calculations for this are as follows:

1. New alpha level at H_1 = old alpha level at H_1 + w_{31} × (the alpha level at H_3) = $\alpha/3 + (1) \times (\alpha/3) = 2\alpha/3$. (The weight w_{31} is for the arrow going from H_3 to H_1.)
2. New alpha level at H_2 = old alpha level at H_2 + w_{32} × (the alpha level at H_3) = $\alpha/3 + (0) \times (\alpha/3) = \alpha/3$. (Note that there is no arrow shown from H_3 to H_2, as its weight $w_{32} = 0$.)
3. New weight w_{12} for the arrow going from H_1 to H_2 = (old w_{12} + A)/(1- B), where A = additional weight for H_1 to H_2 going through H_3 = $w_{13} \times w_{32} = (1/3) \times (0) = 0$, and B = adjustment for the arrow going from H_1 to H_3 and returning back to H_1 = $w_{13} \times w_{31} = (1/3) \times (1) = 1/3$. Therefore, new $w_{12} = (2/3 + 0) /(1 – 1/3) = 1$.
4. Similarly, new weight w_{21} for the arrow going from H_2 to H_1 = (old $w_{21} + w_{23} \times w_{31}$)/(1- $w_{23} \times w_{32}$) = $[1/2 + (1/2) \times (1)]/ [1 - (1/2) \times (0)] = 1$.

This gives the updated graph in Figure A6 (b). Similar calculations can be made for graphs for H_1 and H_3 if H_2 is rejected and for H_2 and H_3 on rejecting H_1.

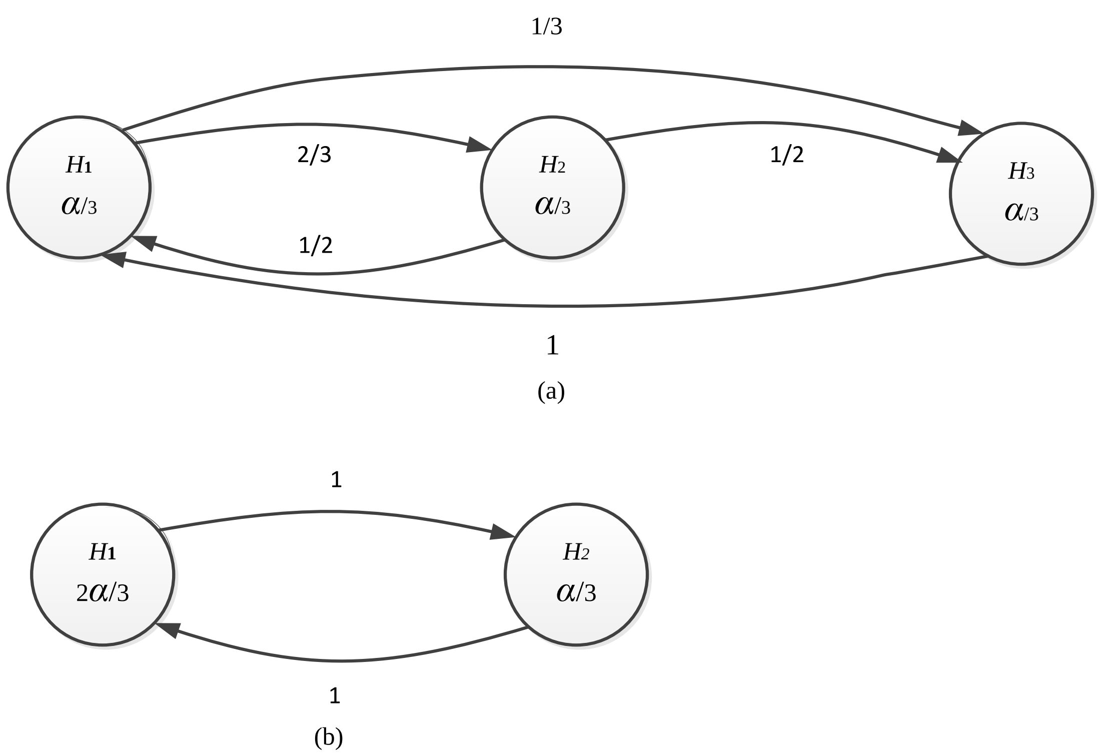


Figure A6: Initial diagram (a) for three hypotheses with fixed weights on the paths connecting the hypotheses, and updated graph (b) when hypotheses H_1 and H_2 are not yet rejected but H_3 is rejected.