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DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.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be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance.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

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

1

Multiple Endpoints in Clinical Trials
Guidance for Industry[1]

4

5

6
7   This draft guidance, when finalized, will represent the current thinking of the Food and Drug
8   Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not
9   binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the
10 applicable statutes and regulations. To discuss an alternative approach, contact the FDA staff responsible
11 for this guidance as listed on the title page.
12

13
14

15 ## I.   INTRODUCTION

16

17 This guidance provides sponsors and review staff with the Agency's thinking about the problems
18 posed by multiple endpoints in the analysis and interpretation of study results and how these
19 problems can be managed in clinical trials for human drugs, including drugs subject to licensing
20 as biological products. Most clinical trials performed in drug development contain multiple
21 endpoints to assess the effects of the drug and to document the ability of the drug to favorably
22 affect one or more disease characteristics. As the number of endpoints analyzed in a single trial
23 increases, the likelihood of making false conclusions about a drug's effects with respect to one or
24 more of those endpoints becomes a concern if there is not appropriate adjustment for
25 multiplicity. The purpose of this guidance is to describe various strategies for grouping and
26 ordering endpoints for analysis and applying some well-recognized statistical methods for
27 managing multiplicity within a study in order to control the chance of making erroneous
28 conclusions about a drug's effects. Basing a conclusion on an analysis where the risk of false
29 conclusions has not been appropriately controlled can lead to false or misleading representations
30 regarding a drug's effects.

31

32 FDA's guidance for industry *E9 Statistical Principles for Clinical Trials* (International Council
33 on Harmonisation E9 guidance, or "ICH E9")[2] is a broad ranging guidance that includes
34 discussion of multiple endpoints. This guidance on multiple endpoints in clinical trials for
35 human drugs provides greater detail on the topic. The issuance of this guidance represents
36 partial fulfillment of an FDA commitment under the Food and Drug Administration
37 Amendments Act (FDAAA) of 2007.
38

---

[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.

39 In general, FDA's guidance documents do not establish legally enforceable responsibilities.
40 Instead, guidances describe the Agency's current thinking on a topic and should be viewed only
41 as recommendations, unless specific regulatory or statutory requirements are cited. The use of
42 the word *should* in Agency guidances means that something is suggested or recommended, but
43 not required.
44
45
46 **II.   BACKGROUND AND SCOPE**
47
48 Failure to account for multiplicity when there are several clinical endpoints evaluated in a study
49 can lead to false conclusions regarding the effects of the drug. The regulatory concern regarding
50 multiplicity arises principally in the evaluation of clinical trials intended to demonstrate
51 effectiveness and support drug approval; however, this issue is important throughout the drug
52 development process.
53
54   **A.   Introduction to Study Endpoints**
55
56 Efficacy endpoints are measures intended to reflect the effects of a drug. They include
57 assessments of clinical events (e.g., mortality, stroke, pulmonary exacerbation, venous
58 thromboembolism), patient symptoms (e.g., pain, dyspnea, depression), measures of function
59 (e.g., ability to walk or exercise), or surrogates of these events or symptoms.
60
61 Because most diseases have more than one consequence, many trials are designed to examine the
62 effect of a drug on more than one endpoint. In some cases, efficacy cannot be adequately
63 established on the basis of a single endpoint. In other cases, an effect on any of several
64 endpoints could be sufficient to support approval of a marketing application. When the rate of
65 occurrence of a single event is expected to be low, it is common to combine several events (e.g.,
66 cardiovascular death, heart attack, and stroke) in a "composite event endpoint" where the
67 occurrence of any of the events would constitute an "endpoint event."
68
69 When there are many endpoints prespecified in a clinical trial, they are usually classified into
70 three families: primary, secondary, and exploratory.

71 • The set of primary endpoints consists of the outcome or outcomes (based on the drug's
72   expected effects) that establish the effectiveness, and/or safety features, of the drug in order
73   to support regulatory action. When there is more than one primary endpoint and success on
74   any one alone could be considered sufficient to demonstrate the drug's effectiveness, the rate
75   of falsely concluding the drug is effective is increased due to multiple comparisons (see
76   section II.E).

77 • Secondary endpoints may be selected to demonstrate additional effects after success on the
78   primary endpoint. For instance, a drug may demonstrate effectiveness on the primary
79   endpoint of survival, after which the data regarding an effect on a secondary endpoint, such
80   as functional status, would be tested. Secondary endpoints may also provide evidence that a
81   particular mechanism underlies a demonstrated clinical effect (e.g., a drug for osteoporosis
82   with fractures as the primary endpoint, and improved bone density as a secondary endpoint).

83 • All other endpoints are referred to as exploratory in this document (see section III.A).
84

85   Endpoints are frequently ordered by clinical importance, with the most important being
86   designated as primary (e.g., mortality or irreversible morbidity). This is not always done,
87   however, for a variety of reasons. The most common reasons not to order endpoints by clinical
88   importance are if there are likely to be too few of the more clinically important endpoint events
89   to provide adequate power for the study, or if the effect on a clinically less important endpoint is
90   expected to be larger. In these cases, endpoints are often ordered by the likelihood of
91   demonstrating an effect. For example, time-to-disease progression is often selected as the
92   primary endpoint in oncology trials even though survival is almost always the most important
93   endpoint; the reasons being that an effect on disease progression may be more readily
94   demonstrable, may be detected earlier, and often has a larger effect size because the observed
95   effect on survival can be diluted by subsequent treatment post-progression. Section III.A
96   includes further discussion of the primary and secondary endpoint families. The determination
97   of which endpoints are primary, secondary, or exploratory, regardless of the reasons for the
98   determination, should always be made prospectively (see ICH E9).
99
100   Although this guidance focuses on endpoints intended to demonstrate effectiveness, a study that
101   is designed specifically to assess safety outcomes may also have both primary and secondary
102   endpoints, which would then be subject to the same multiplicity considerations described in this
103   guidance.
104

105   **B.**   **Demonstrating the Study Objective of Effectiveness**
106

107   A conclusion that a study has demonstrated an effect of a drug is critical to meeting the legal
108   standard for substantial evidence of effectiveness required to support approval of a new drug
109   (i.e., "… adequate and well-controlled investigations…on the basis of which it could fairly and
110   responsibly be concluded…that the drug will have the effect it purports…to have…") (section
111   505(d) of the FD&C Act).[3] FDA regulations further establish that to be adequate and well
112   controlled, a clinical study of a drug must include, among other things, "an analysis of the results
113   of the study adequate to assess the effects of the drug," a requirement that furthers the "purpose
114   of conducting clinical investigations of a drug" which is "to distinguish the effect of a drug from
115   other influences, such as spontaneous change in the course of the disease, placebo effect, or
116   biased observation."[4] The clinical trial community has accepted an approach that finds a
117   treatment effect to be established when a determination is made that the apparent treatment effect
118   observed in a clinical trial is not likely to have occurred by chance. This is generally
119   accomplished by placing a limit on the probability that the finding is the result of chance.
120

---

[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).

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

126   •   One hypothesis, the *null hypothesis*, states that there is no treatment effect on the chosen
127   clinical endpoint. The treatment effect is represented by a parameter, for example, T-C,
128   the difference between the test group's average outcome measure (T) and that of the
129   control group (C), or T/C, the ratio of response rates for the two groups. The null
130   hypothesis is represented by the equation T-C = 0 or T/C = 1, stating that the true
131   difference between the outcomes for the test group and the control group is zero or the
132   risk ratio is 1 (i.e., there is no treatment effect).

133   •   The other hypothesis is called the *alternative hypothesis* and posits that there is at least
134   some treatment effect of the test drug, usually represented as T-C > 0 (or T/C >1) for the
135   alternative of interest (a beneficial effect of the drug).
136
137   The *test of hypothesis* determines whether (1) the trial results are consistent with the null
138   hypothesis of no treatment effect or (2) the favorable result of the trial is so unlikely to have been
139   obtained if the null hypothesis were true that the null hypothesis can be rejected and the
140   alternative hypothesis, that there is a treatment effect, accepted.
141
142   Sometimes (e.g., in some vaccine trials), demonstration of an effect of at least some minimum
143   size is considered essential for approval of a drug. In this case the null hypothesis might be
144   modified to T-C $\leq$ m or T/C $\leq$ r, where m or r is the smallest effect that could be accepted. Such
145   modifications of the null hypothesis can have an impact on the sample size of a trial.
146
147   **C.**   **Type I Error**
148
149   The rejection of the null hypothesis supports the study conclusion that there is a difference
150   between treatment groups but does not constitute absolute proof that the null hypothesis is false.
151   There is always some possibility of mistakenly rejecting the null hypothesis when it is, in fact,
152   true. Such an erroneous conclusion is called a Type I error. Null hypothesis rejection is based
153   on a determination that the probability of observing a result at least as extreme as the result of the
154   study assuming the null hypothesis is true (the p-value) is sufficiently low. The probability of
155   concluding that there was a difference between treatment groups due to the drug when, in fact,
156   there was none, is called the Type I error probability or rate, denoted as alpha (α).
157
158   Type I error probabilities can apply to two-sided hypothesis tests, in which case they refer to the
159   probability of concluding that there is a difference (beneficial or harmful) between the drug and
160   control when there is no difference. Type I error probabilities can also apply to one-sided
161   hypothesis tests, in which case they refer to the probability of concluding specifically that there
162   is a *beneficial difference* due to the drug when there is not. The most widely-used values for
163   alpha are 0.05 for two-sided tests and 0.025 for one-sided tests. In the case of one-sided tests, an
164   alpha of 0.025 means that the probability of falsely concluding a beneficial effect of the drug
165   when none exists is no more than 2.5 percent, or 1 chance in 40 (represented as p $\leq$ 0.025). In
166   the case of two-sided tests, an alpha of 0.05 means that the probability of falsely concluding that

167   the drug differs from the control in either direction (benefit or harm) when no difference exists is
168   no more than 5 percent, or 1 chance in 20 (represented as $p \le 0.05$). Use of a two-sided test with
169   an alpha of 0.05 generally also ensures that the probability of falsely concluding *benefit* when
170   there is none is no more than approximately 2.5 percent (1 chance in 40). Use of either test
171   therefore provides strong assurance against the possibility of a false-positive result (i.e., no more
172   than 1 chance in 40) and a sound basis for regulatory decision-making, especially when
173   substantiated by another study or other confirmatory evidence.[5]

174
175   For simplicity, this guidance discusses statistical testing of two-sided hypotheses at the 5 percent
176   level, with the understanding that the one-sided alternative hypothesis of a beneficial drug effect
177   is our focus, and the chance of a false positive conclusion is our primary concern. In most cases,
178   sponsors can perform either two-sided or one-sided tests of hypothesis, at their discretion.

179
180   This discussion is focused on the study's final analysis. If interim analyses occur during a study,
181   there should be a prospective plan to ensure that these additional analyses do not increase the
182   chances of a false positive conclusion. When multiple endpoints are examined at an interim
183   analysis, the appropriate adjustments can become complex; discussion of this issue is outside the
184   scope of this guidance.

185
186   FDA's concern for controlling the Type I error probability is to minimize the chances of a false
187   favorable conclusion for any of the primary or secondary endpoints, regardless of which and how
188   many endpoints in the study have no effect (called strong control of the Type I error probability).
189   Determining if strong control is achieved can be complicated when more than one endpoint is
190   under consideration, any one of which could support a conclusion that the treatment has a
191   beneficial effect. When there is more than one study endpoint, care must be taken to ensure that
192   the evaluation of multiple hypotheses does not lead to inflation of the study's overall Type I error
193   probability, called the study-wise Type I error probability, which is the chance of a false positive
194   conclusion on any planned endpoint analysis.

195
196   The discussion of specific statistical methods for managing multiplicity in section IV illustrates
197   that when some of the null hypotheses should be rejected but others should not be rejected, the
198   control of the Type I error probability can become complex. The challenge that arises from
199   testing multiple hypotheses associated with multiple endpoints in a study is to ensure that there is
200   a proper accounting for all of the possible ways the endpoints of the study could produce false
201   positive conclusions (see section II.E).

202
203   An essential element of Type I error rate control is the prospective specification of:
204   • all endpoints that will be tested and
205   • all data analyses that will be performed to test hypotheses about the prespecified
206   endpoints.
207   For a multiple endpoints study, the analysis plan should describe how (or ways to determine
208   how) the endpoints are tested, including the order of testing and the alpha level applied to each
209   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.

210
211   **D.**   **Relationship Between the Observed and True Treatment Effects**
212
213 The statistical analysis associated with a hypothesis test produces three primary measures of
214 interest:
215   • a point estimate,
216   • a confidence interval, and
217   • a p-value.
218 The effect of the treatment is typically presented as a point estimate (the observed T-C
219 difference) that represents the most likely true effect. The confidence interval is usually two-
220 sided and illustrates the range of true treatment effect values consistent with the data observed in
221 the trial.
222
223 In addition to the point estimate of the treatment effect, it is important to consider the width of
224 the confidence interval. The confidence interval provides a measure of the precision of the
225 estimate of the treatment effect. The narrower the confidence interval, and the further away its
226 lower bound is from the null hypothesis of no treatment effect (T-C = 0 or T/C = 1), the more
227 confident we are of both the magnitude and existence of the treatment effect. Generally, the
228 farther the lower bound of the confidence interval is from zero (or 1), the more persuasive
229 (smaller) the p-value is and the lower the likelihood that the effectiveness finding was a chance
230 occurrence.
231
232 There is usually a relationship between the test of a hypothesis and the confidence interval; each
233 focuses on related but not identical questions:
234   • The test of a hypothesis focuses on whether or not there is an effect.
235   • The confidence interval focuses on the magnitude of the effect and the precision with
236   which we know it.
237 The emphasis of this guidance is not on the confidence interval, but rather on the test of a
238 hypothesis, where the issue is whether a treatment effect on a particular endpoint exists at all.
239 Although confidence intervals are also critical to the interpretation of an effect when one exists,
240 determining the confidence interval with some of the statistical methods for managing
241 multiplicity described in section IV is complex. The primary goal of this guidance is to provide
242 recommendations for designing studies that control the chances of erroneously concluding that a
243 treatment is effective with respect to a particular endpoint. In some areas, however, confidence
244 intervals are used to test hypotheses of the type described at the end of section II.B (e.g., T-C $\leq$
245 m). In these situations, it is critical to ensure that the confidence intervals appropriately reflect
246 multiplicity of hypothesis tests.
247
248   **E.**   **Multiplicity**
249
250 As described in section I.A, clinical trials often include more than one endpoint as an indicator of
251 effectiveness. When a trial is designed so that more than one study endpoint or comparison (of
252 treatment to control) could lead to a conclusion that effectiveness was established, testing each
253 endpoint separately at $\alpha = 0.05$ will inflate the Type I error rate and overstate the statistical
254 significance. The inflation of the Type I error rate can be quite substantial if there are many

255   comparisons. Because this form of Type I error rate inflation is the result of multiple
256   comparisons, it is termed a multiplicity problem.
257
258   In a clinical trial with a single endpoint tested at $\alpha = 0.05$, the probability of finding a difference
259   between the treatment group and a control group by chance alone is at most 0.05 (a 5 percent
260   chance). By contrast, if there are two independent endpoints, each tested at $\alpha = 0.05$, and if
261   success on either endpoint by itself would lead to a conclusion of a drug effect, there is a
262   multiplicity problem. For each endpoint individually, there is at most a 5 percent chance of
263   finding a treatment effect when there is no effect on the endpoint, and the chance of erroneously
264   finding a treatment effect on at least one of the endpoints (a false positive finding) is about 10
265   percent. More precisely, when the endpoints are independent, there is a 95 percent chance of
266   correctly failing to detect an effect for each endpoint if there is no true effect for either endpoint.
267   The chance of correctly failing to detect an effect on both endpoints together is thus 0.95 * 0.95,
268   which equals 0.9025, and so the probability of falsely detecting an effect on at least one endpoint
269   is 1 - 0.9025, which equals 0.0975. Without correction, the chance of making a Type I error for
270   the study as a whole would be 0.1 and the study-wise Type I error rate is therefore not
271   adequately controlled. The problem is exacerbated when more than two endpoints are
272   considered. For three endpoints, the Type I error rate is 1 - (.95 * .95 * .95), which is about 14
273   percent. For ten endpoints, the Type I error rate is about 40 percent.
274
275   Even when a single outcome variable is being assessed, if the approach to evaluating the study
276   data is to analyze multiple facets of that outcome (e.g., multiple dose groups, multiple time
277   points, or multiple patient subgroups based on demographic or other characteristics) and regard
278   the study as positive (i.e., conclude that the drug has been shown to produce a beneficial effect)
279   if any one analysis is positive, the multiplicity of analyses causes inflation of the Type I error
280   rate, thus increasing the probability of reaching a false conclusion about the effects of the drug.
281   Similarly, application of more than one analytic approach to one endpoint introduces multiplicity
282   by providing additional ways for the trial to be successful (to "win"). Examples include
283   conducting both unadjusted and covariate-adjusted analyses, use of different analysis populations
284   (intent-to-treat, completers, per protocol), use of different endpoint assessments (by investigator
285   vs. a central endpoint assessment committee), and many others. By inflating Type I error,
286   multiplicity produces uncertainty in interpretation of the study results such that the strength of a
287   finding becomes unclear, and conclusions about whether effectiveness has been demonstrated in
288   the study become unreliable. There are various approaches that can be planned prospectively
289   and applied to maintain the Type I error rate at 5 percent. Among these are adjustments to the
290   alpha level for determining that an individual endpoint test is positive, structuring the order in
291   which the endpoints are tested, and others. These approaches are discussed in detail in section
292   IV.
293
294   An important principle for controlling multiplicity is to prospectively specify all planned
295   endpoints, time points, analysis populations, and analyses. Once these factors are specified,
296   appropriate adjustments for multiple endpoints and analyses can be planned and applied, as
297   needed. Changes in the analytic plan to perform additional analyses, however, can reintroduce a
298   multiplicity problem that can negatively impact the ability to interpret the study's results unless
299   these changes are made prior to data analysis and appropriate multiplicity adjustments are
300   performed. In the past, it was not uncommon, after the study was unblinded and analyzed, to see

301   a variety of post hoc adjustments of design features (e.g., endpoints, analyses), usually plausible
302   on their face, to attempt to elicit a positive study result from a failed study — a practice
303   sometimes referred to as data-dredging. Although post hoc analyses of trials that fail on their
304   prospectively specified endpoints may be useful for generating hypotheses for future testing,
305   they do not yield definitive results. The results of such analyses can be biased because the
306   choice of analyses can be influenced by a desire for success. The results also represent a
307   multiplicity problem because there is no way to know how many different analyses were
308   performed and there is no credible way to correct for the multiplicity of statistical analyses and
309   control the Type I error rate. Consequently, post hoc analyses by themselves cannot establish
310   effectiveness. Also, additional endpoints that have not been pre-specified or evaluated with
311   adjustment for multiplicity when required cannot, in general, be used to demonstrate an effect of
312   the drug, even in successful studies.

313

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

326

327   The focus of this guidance is control of the Type I error rate for the planned primary and
328   secondary endpoints of a clinical trial so that the major findings are well supported and the
329   effects of the drug have been demonstrated. Once a trial is successful (demonstrates
330   effectiveness or "wins" on the primary endpoint(s)), there are many other attributes of a drug's
331   effects that may be described. Analyses that describe these other attributes of a drug can be
332   informative and are often included in physician labeling.[6] Examples include: the time course of
333   treatment effects;[7] the full distribution of responses amongst participants;[8] treatment effects on
334   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.

335   Nevertheless, it is important to understand that these descriptions with respect to additional
336   attributes are not demonstrated additional effects of a drug unless the analyses were prespecified,
337   and appropriate multiplicity adjustments were applied. Therefore, presenting p-values from
338   descriptive analyses (that is, from analyses that were not prespecified and for which appropriate
339   multiplicity adjustments were not applied) is inappropriate because doing so would imply a
340   statistically rigorous conclusion and convey a level of certainty about the effects that is not
341   supported by that trial. Descriptive analyses are not the subject of this guidance and are not
342   addressed in detail.
343
344   In the following sections, the issues of multiple endpoints and methods to address them are
345   illustrated with examples of different study endpoints. Both the issues and methods that apply to
346   multiple endpoints also apply to other sources of multiplicity, including multiple doses, time
347   points, or study population subgroups.
348
349   **III.   MULTIPLE ENDPOINTS: GENERAL PRINCIPLES**
350
351   **A.   The Hierarchy of Families of Endpoints**
352
353   Endpoints in adequate and well-controlled drug trials are usually grouped hierarchically, often
354   according to their clinical importance, but also taking into consideration the expected frequency
355   of the endpoint events and anticipated drug effects. The critical determination for grouping
356   endpoints is whether they are intended to establish effectiveness to support approval or intended
357   to demonstrate additional meaningful effects. Endpoints essential to establish effectiveness for
358   approval are called *primary endpoints*. *Secondary endpoints* may be used to support the primary
359   endpoint(s) and/or demonstrate additional effects. The third category in the hierarchy includes
360   all other endpoints, which are referred to as exploratory. *Exploratory endpoints* may include
361   clinically important events that are expected to occur too infrequently to show a treatment effect
362   or endpoints that for other reasons are thought to be less likely to show an effect but are included
363   to explore new hypotheses. Each category in the hierarchy may contain a single endpoint or a
364   family of endpoints.
365
366   *1. Primary Endpoint Family*
367
368   The endpoint(s) that will be the basis for concluding that the study met its objective (i.e., the
369   study "wins") is designated the primary endpoint or primary endpoint family. When there is a
370   single pre-specified primary endpoint, there are no multiple endpoint-related multiplicity issues
371   in the determination that the study achieved its objective; however, there could still be
372   multiplicity issues for demonstration of effects on secondary endpoints.
373
374   Multiple primary endpoints occur in three ways, further described in section III.C. The first is
375   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.

376 this case, failure to adjust for multiplicity can lead to a false conclusion that the drug is effective.
377 The second is where determination of effectiveness depends on success on all of two or more
378 primary endpoints. In this setting, there are no multiple endpoint-related multiplicity issues, and
379 therefore, no concern with Type I error rate inflation, but there is a concern with Type II error
380 rate inflation (See Section III.B). In the third, critical aspects of effectiveness can be combined
381 into a single primary composite (or other multicomponent) endpoint, thereby avoiding multiple
382 endpoint-related multiplicity issues. For example, in many cardiovascular studies it is usual to
383 combine several endpoints (e.g., cardiovascular death, heart attack, and stroke) into a single
384 composite endpoint that is primary and to consider death a secondary endpoint (section III.A.2).
385 A comprehensive examination of the drug's effects earlier in development might aid in the
386 selection of a sensitive and informative measure of the drug's effect and allow use of a single
387 primary endpoint for the confirmatory trial.

389   2.   *Secondary Endpoint Family*

391 The collection of all secondary endpoints is called the secondary endpoint family. Secondary
392 endpoints are those that may provide supportive information about a drug's effect on the primary
393 endpoint or demonstrate additional effects on the disease or condition. Secondary endpoints
394 might include a pharmacodynamic effect that would not be considered an acceptable primary
395 efficacy endpoint but is closely related to the primary endpoint, (e.g., an effect consistent with
396 the drug's purported mechanism of action). A secondary endpoint could be a clinical effect
397 related to the primary endpoint that extends the understanding of that effect (e.g., an effect on
398 survival when a cardiovascular drug has shown an effect on the primary endpoint of heart
399 failure-related hospitalizations) or provide evidence of a clinical benefit distinct from the effect
400 shown by the primary endpoint (e.g., a disability endpoint in a multiple sclerosis treatment trial
401 in which relapse rate is the primary endpoint). In all cases, when an effect on the primary
402 endpoint is shown, the secondary endpoints can be examined and may contribute important
403 supportive information about a drug's effectiveness.

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

419 Because study sample size is often determined based only on the amount of information needed
420 to adequately assess the primary hypothesis, many studies lack sufficient power to demonstrate

421 effects on secondary endpoints. If success on the secondary endpoints is important, the
422 secondary endpoints should be considered when determining study design (e.g., sample size).
423
424 An example of a secondary endpoint used to further characterize the drug's effect is a
425 measurement of the primary outcome variable at 30 days in a trial whose primary endpoint is the
426 same outcome measured at 6 months. Another example is a secondary endpoint of the
427 percentage of patients whose symptoms are "very improved," when the primary endpoint is the
428 percentage of patients with any amount of improvement for the same symptoms. Adjustment for
429 multiplicity is necessary to demonstrate these additional effects.
430
431 It is recommended that the list of secondary endpoints be short, because the chance of
432 demonstrating an effect on any secondary endpoint after appropriate correction for multiplicity
433 becomes increasingly small as the number of endpoints increases. Endpoints intended to serve
434 the purpose of hypothesis generation should not be included in the secondary endpoint family.
435 These should be considered exploratory endpoints.
436
437 **B.   Type II Error Rate and Multiple Endpoints**
438
439 One of the greatest concerns in the design of clinical trials intended to support drug approval is
440 inflation of the Type I error rate, because it can lead to an erroneous conclusion that a drug is
441 effective. FDA is also concerned with the risk of Type II error, which is failing to show an effect
442 of a drug where there actually is one. The intended level of risk of a Type II error is usually
443 denoted by the symbol beta *(β)*. The study's likelihood of avoiding Type II error (1-*β*), if the
444 drug actually has the specified effect, is called study power. The desired power is an important
445 factor in determining the sample size.
446
447 The sample size of a study is generally chosen to provide a reasonably high power to show a
448 treatment effect if an effect of a specified size is in fact present. In addition to the treatment
449 effect, the optimal sample size of a study is influenced by the variability of the endpoint and the
450 alpha level specified for the test of hypothesis for that endpoint. Investigators should consider
451 these factors for all of the endpoints for which the study is intended to be well powered.
452
453 Many of the statistical adjustment methods to control the Type I error rate for multiplicity
454 discussed in section IV decrease study power because they lower the alpha level used for each of
455 the individual endpoints' test of hypothesis, making it more difficult to achieve statistical
456 significance. Increasing the sample size appropriately can overcome this decrease in power. In
457 general, the greater the number of endpoints (analyses), the greater the statistical adjustment that
458 is needed and the greater the increase in the sample size of the trial necessary to maintain power
459 for all individual endpoints. This decrease in study power (i.e., increased Type II error rate)
460 from multiplicity is often a practical limiting factor in choosing the number of endpoints
461 designated for a trial as indicators of success without requiring an excessive sample size.
462
463 Some of the methods discussed in section IV to manage multiplicity are complex and may, for
464 example, call for the alpha level for any particular test of hypothesis to be determined by the
465 actual study endpoint results and the resulting sequence of hypothesis testing. In some cases,
466 sponsors may wish to have the study well powered for one or two secondary endpoints in

467 addition to the primary endpoint family, further adding to the complexity. Determination of an
468 appropriate study sample size to ensure that the study is appropriately powered can be difficult in
469 these cases, and often will be dependent upon computer simulations rather than an analytic
470 formula, which can be used for simpler situations.
471
472 The use of two or more endpoints for which demonstration of an effect on each is needed to
473 support regulatory approval (called co-primary endpoints; see section III.C.1 below) increases
474 the Type II error rate and decreases study power. If, for example, the study sample size is
475 selected to provide 80 percent power to show success on each of two endpoints (i.e., Type II
476 error rate is 20 percent for each), and the endpoints are entirely independent, the power to show
477 success on both will be just 64 percent (0.8 x 0.8): i.e., the likelihood of the study failing to
478 support a conclusion of a favorable drug effect when such an effect existed (the Type II error
479 rate) would be 36 percent. The study power could, of course, be restored by increasing the
480 sample size. Multiplicity and Type I error rate inflation are not a concern with co-primary
481 endpoints, as there is only one way to succeed.
482
483 The loss of power may not be so severe when the endpoints are correlated (i.e., not fully
484 independent). With positive correlation, there is an increased chance that a second endpoint will
485 demonstrate the treatment effect if one endpoint is successful, potentially increasing study power
486 well above the 64 percent estimate. Moreover, the individual endpoints usually do not all have
487 the same power-influencing characteristics because the effect size and variability estimates may
488 be different for the different endpoints. If the study is designed so that a test of the endpoint
489 upon which it is most difficult to demonstrate an effect has 80 percent power, the other endpoints
490 may have power in excess of 80 percent to show the expected effect. In that case, the overall
491 study power, even if the endpoints were fully independent, will also be higher than if all
492 endpoints were equally powered. Nonetheless, when considering use of co-primary endpoints in
493 a study, it should be recognized that use of more than two can markedly reduce study power.
494

495   **C.   Types of Multiple Endpoints**
496
497 Multiple endpoints may be needed when determining that the drug confers a clinical benefit
498 depends on more than one disease aspect or outcome being affected. Multiple endpoints may
499 also be used when (1) there are several important aspects of a disease or several ways to assess
500 an important aspect, (2) there is no consensus about which one will best serve the study
501 purposes, and (3) an effect on any one will be sufficient as evidence of effectiveness to support
502 approval. In some cases, multiple aspects of a disease may appropriately be combined into a
503 single endpoint, but subsequent analysis of the components is generally important for an
504 adequate understanding of the drug's effect. These circumstances when multiple endpoints are
505 encountered are discussed below.
506
507   *1.   When Demonstration of Treatment Effects on All of Two or More Distinct*
508   *Endpoints Is Necessary to Establish Clinical Benefit (Co-Primary Endpoints)*
509
510 The primary endpoint for determining that a drug is effective should encompass one or more of
511 the important features of a disorder and should be clinically meaningful. There are two types of
512 circumstances when no single endpoint adequately serves this purpose.

513
514 For some disorders, there are two or more different features that are so critically important to the
515 disease under study that a drug will not be considered effective without demonstration of a
516 treatment effect on all of these disease features. The term used in this guidance to describe this
517 circumstance of multiple primary endpoints is co-primary endpoints. Multiple primary endpoints
518 become co-primary endpoints when it is necessary to demonstrate an effect on each of the
519 endpoints to conclude that a drug is effective.
520
521 Therapies for the treatment of migraine headaches illustrate this circumstance. Although pain is
522 the most prominent feature, migraine headaches are also often characterized by the presence of
523 photophobia, phonophobia, and nausea, all of which are clinically important. Which of the three
524 is most clinically important varies among patients. A recent approach to studying treatments is
525 to consider a drug effective for migraines only if pain and an individually-specified most
526 bothersome second feature are both shown to be improved by the drug treatment.
527
528 A second kind of circumstance in which a demonstration of an effect on two endpoints is needed
529 is when there is a single identified critical feature of the disorder, but uncertainty as to whether
530 an effect on the endpoint alone is clinically meaningful. In these cases, two endpoints are often
531 used. One endpoint is specific for the disease feature intended to be affected by the drug but not
532 readily interpretable as to the clinical meaning, and the second endpoint is clinically interpretable
533 but may be less specific for the intended action of the test drug. A demonstration of
534 effectiveness is dependent upon both endpoints showing a drug effect. One endpoint ensures the
535 effect occurs on the core disease feature, and the other ensures that the effect is clinically
536 meaningful.
537
538 An example illustrating this second circumstance is development of drugs for treatment of the
539 symptoms of Alzheimer's disease. Drugs for Alzheimer's disease have generally been expected
540 to show an effect on both the defining feature of the disease, decreased cognitive function, and
541 on some measure of the clinical impact of that effect. Because there is no single endpoint able to
542 provide convincing evidence of both, co-primary endpoints are used. One primary endpoint is
543 the effect on a measure of cognition in Alzheimer's disease (e.g., the Alzheimer's Disease
544 Assessment Scale-Cognitive Component), and the second is the effect on a clinically
545 interpretable measure of function, such as a clinician's global assessment or an Activities of
546 Daily Living Assessment.
547
548 Trials of combination vaccines are another situation in which co-primary endpoints are
549 applicable. These vaccine trials are typically designed and powered for demonstration of a
550 successful outcome on effectiveness endpoints for each pathogen against which the vaccine is
551 intended to provide protection.
552
553 As discussed in section II.E, multiplicity problems occur when there is more than one way to
554 determine that the study is a success. When using co-primary endpoints, however, there is only
555 one result that is considered a study success, namely, that all of the separate endpoints are
556 statistically significant. Therefore, testing all of the individual endpoints at the 0.05 level does
557 not cause inflation of the Type I error rate; rather, the impact of co-primary endpoint testing is to
558 increase the Type II error rate. The size of this increase will depend on the correlation of the co-

559 primary endpoints. In general, unless clinically very important, the use of more than two co-
560 primary endpoints should be carefully considered because of the loss of power.
561
562 There have been suggestions that the statistical testing criteria for each co-primary endpoint
563 could be relaxed (e.g., testing at an alpha of 0.06 or 0.07) to accommodate the loss in statistical
564 power arising from the need to show an effect on both endpoints. Relaxation of alpha is
565 generally not acceptable because doing so would undermine the assurance of an effect on each
566 disease aspect considered essential to showing that the drug is effective in support of approval.
567
568   *2.*   *When Demonstration of a Treatment Effect on at Least One of Several Primary*
569   *Endpoints Is Sufficient*
570
571 Many diseases have multiple sequelae, and an effect demonstrated on any one of these aspects
572 may support a conclusion of effectiveness. Selection of a single primary endpoint may be
573 difficult, however, if the aspect of a disease that will be responsive to the drug or the evaluation
574 method that will better detect a drug effect is not known a priori (at the time of trial design). In
575 this circumstance, a study might be designed such that success on any one of several endpoints
576 could support a conclusion of effectiveness. This creates a primary endpoint family. For
577 example, consider a drug for the treatment of burn wounds where it is not known whether the
578 drug will increase the rate of wound closure or reduce scarring, but the demonstration of either
579 effect alone would be considered to be clinically important. A study in this case might have both
580 wound closure rate and a scarring measure as separate primary endpoints.
581
582 This use of multiple endpoints creates a multiplicity problem because there are several ways for
583 the study to successfully demonstrate a treatment effect. Control of the Type I error rate for the
584 primary endpoint family is critical. A variety of approaches can be used to address this
585 multiplicity problem; section IV is devoted to describing and discussing some of these
586 approaches.
587
588 It should be noted that failure to demonstrate an effect on any one of the individual prespecified
589 primary endpoints does not preclude making valid conclusions with respect to the other
590 prespecified primary endpoints. From a regulatory perspective, the results for all of the
591 prespecified primary endpoints, both positive and negative, are considered in the overall
592 assessment of risks and benefits.
593
594   *3.*   *Composite Endpoints*
595
596 There are some disorders for which more than one clinical outcome in a clinical trial is
597 important, and all outcomes are expected to be affected by the treatment. Rather than using each
598 as a separate primary endpoint (creating multiplicity) or selecting just one to be the primary
599 endpoint and designating the others as secondary endpoints, it may be appropriate to combine
600 those clinical outcomes into a single variable. This is called a "composite endpoint," where an
601 endpoint is defined as the occurrence or realization in a patient of any one of the specified
602 components. When the components correspond to distinct events, composite endpoints are often
603 assessed as the time to first occurrence of any one of the components, but in diseases where a
604 patient might have more than one event, it also may be possible to analyze total endpoint events

605   (see section III.D.1). A single statistical test is performed on the composite endpoint;
606   consequently, no multiplicity problem occurs and no statistical adjustment is needed.
607
608   An important reason for using a composite endpoint is that the incidence rate of each of the
609   events may be too low to allow a study of reasonable size to have adequate power; the composite
610   endpoint can provide a substantially higher overall event rate that allows a study with a
611   reasonable sample size and study duration to have adequate power. Composite endpoints are
612   often used when the goal of treatment is to prevent or delay morbid, clinically important but
613   uncommon events (e.g., use of an anti-platelet drug in patients with coronary artery disease to
614   prevent myocardial infarction, stroke, and death).
615
616   The choice of the components of a composite endpoint should be made carefully. Because the
617   occurrence of any one of the individual components is considered to be an endpoint event, each
618   of the components is of equal importance in the analysis of the composite. The treatment effect
619   on the composite rate can be interpreted as characterizing the overall clinical effect when the
620   individual events all have reasonably similar clinical importance. The effect on the composite
621   endpoint, however, will not be a reasonable indicator of the effect on all of the components or an
622   accurate description of the drug's benefit, if the clinical importance of different components is
623   substantially different and the drug effect is chiefly on the least important event. Furthermore, it
624   is possible that a component with greater importance may appear to be adversely affected by the
625   treatment, even if one or more event types of lesser importance are favorably affected, so that
626   although the overall outcome still has a favorable statistical result, doubt may arise about the
627   treatment's clinical value. In this case, although the overall statistical analysis indicates the
628   treatment is successful, careful examination of the data may call this conclusion into question.
629   For this reason, as well as for a greater depth of understanding of the treatment's effects,
630   analyses of the components of the composite endpoint are important (see section III.D) and can
631   influence interpretation of the overall study results.
632
633   *4.   Other Multi-Component Endpoints*
634
635   A different type of multi-component endpoint is a within-patient combination of two or more
636   components. In this type of endpoint, an individual patient's evaluation is dependent upon
637   observation of all of the specified components in that patient. A single overall rating or status is
638   determined according to specified rules.
639
640   When the components are ordered categorical or continuous numeric scales, one way of forming
641   an overall rating is to use the sum or average across the individual domain scores. Study
642   hypotheses are then tested by comparing the overall mean values between groups. Examples of
643   this type are the Positive and Negative Syndrome Scale (PANSS) in schizophrenia research; the
644   Toronto Western Spasmodic Torticollis Rating Scale for evaluating cervical dystonia; the
645   Hamilton Rating Scale for Depression (HAM-D); the Brief Psychiatric Rating Scale; and many
646   patient-reported outcomes (PROs).
647
648   Alternatively, a multi-component endpoint may be a dichotomous (event) endpoint
649   corresponding to an individual patient achieving specified criteria on each of the multiple
650   components. This dichotomous form of a multi-component endpoint might be preferred over

651 multiple independent endpoints in conditions where assuring individual patients have benefit on
652 all of several disease features is important. For example, the FDA guidance for industry
653 *Considerations for Allogeneic Pancreatic Islet Cell Products* recommends that the primary
654 endpoint in clinical trials of allogeneic pancreatic islet cells for Type 1 diabetes mellitus be a
655 composite in which patients are considered responders only if they meet two dichotomous
656 response criteria: normal range of HbA1c and elimination of hypoglycemia. In contrast, when
657 separate endpoints are analyzed as co-primary endpoints (i.e., all of the several identified disease
658 aspects are required to show an effect), the study provides evidence that the drug affects all of
659 the endpoints on a group-wise basis, but does not ensure an increase in the number of individual
660 patients for whom all endpoints are favorably affected.
661
662 More complex endpoint formulations may be appropriate when there are several different
663 features of a disease that are important, but not all features must be positively affected for a
664 patient to be regarded as receiving benefit. For example, a positive response for an individual
665 patient might be defined as improvement in one or two specific required aspects of a disease
666 along with improvement in at least one, but not all, identified additional disease features, as in
667 the American College of Rheumatology (ACR) scoring system for rheumatoid arthritis. The
668 ACR20 criteria for defining a response to treatment are a 20 percent improvement in two specific
669 disease features (tender joints and swollen joints) and a 20 percent improvement in at least three
670 of five additional features (pain, acute phase reactants, global assessment by patient or physician,
671 or disability). Generally, these types of endpoints are very disease-specific, and clinical research
672 on the particular disease and its manifestations guides the development of such defined, complex
673 combinations of assessments. These combinations, despite incorporating multiple different
674 features of the disease, provide a single primary endpoint for evaluating efficacy and do not raise
675 multiplicity concerns.
676
677 The use of within-patient multi-component endpoints can be efficient if the treatment effects on
678 the different components are generally concordant. Study power can be adversely affected,
679 however, if there is limited correlation among the endpoints. Although multi-component
680 endpoints can provide some gains in efficiency compared to co-primary endpoints, the
681 appropriateness of a particular within-patient multi-component endpoint is generally determined
682 by clinical, rather than statistical, considerations. Formal statistical analyses of these
683 components without prespecification and adjustment for multiplicity, however, may lead to a
684 false conclusion about the effects of the drug with respect to each individual component, as
685 discussed in section III.D.
686
687   *5. Clinically Critical Endpoints Too Infrequent for Use as a Primary Endpoint*
688
689 For many serious diseases, there is an endpoint of such great clinical importance that it is
690 unreasonable not to collect and analyze the endpoint data; the usual example is mortality or
691 major morbidity events (e.g., stroke, fracture, pulmonary exacerbation). Even if relatively few of
692 these events are expected to occur in the trial, they may be included in a composite endpoint (see
693 section III.C.3) and also designated as a planned secondary endpoint to potentially support a
694 conclusion regarding effect on that separate clinical endpoint, if the effect of the drug on the
695 composite primary endpoint is demonstrated. There have been situations, however, where the
696 effect on the primary endpoint was not found to be statistically significant, but there did appear

697 to be an effect on mortality or major morbidity. In the absence of a demonstrated treatment
698 effect on the primary endpoint, secondary endpoints cannot be assessed statistically, but the
699 suggestion of a favorable result on a major outcome such as mortality may be difficult to ignore.
700
701 One approach to avoid this situation would be to designate the mortality or morbidity endpoint as
702 another primary endpoint, and apply one of the statistical methods of section IV with unequal
703 splitting of the alpha. In this way, the endpoint can be validly tested, and should the effect be
704 large, it will provide evidence of efficacy. Depending upon how alpha is allocated, the increase
705 in sample size to maintain study power may only be modest.
706
707 **D.   The Individual Components of Composite and Other Multi-Component**
708 **Endpoints**
709
710 *1.   Evaluating the Components of Composite Endpoints*
711
712 For composite endpoints whose components correspond to events, an event is usually defined as
713 the first occurrence of any of the designated component events. Such composites can be
714 analyzed either with comparisons of proportions between study groups at the end of the study or
715 using time-to-event analyses. The time-to-event method of analysis is the more common method
716 when, within the study's timeframe of observation, the duration of being event-free is clinically
717 meaningful. Although there is an expectation that the drug will have a favorable effect on all the
718 components of a composite endpoint, that is not a certainty. Results for each component event
719 should therefore be individually examined and should always be included in study reports.
720 These analyses will not alter a conclusion about the statistical significance of the composite
721 primary endpoint and are considered descriptive analyses, not tests of hypotheses. If there is,
722 however, an interest in analyzing one or more of the components of a composite endpoint as
723 distinct hypotheses to demonstrate effects of the drug, the hypotheses should be part of the
724 prospectively specified statistical analysis plan that accounts for the multiplicity this analysis will
725 entail, as described above for mortality.
726
727 In analyzing the contribution of each component of a composite endpoint, there are two
728 approaches that differ in how patients who experience more than one of the event-types are
729 considered.
730   • One approach considers only the initial event in each patient. This method displays the
731   incidence of each type of component event only when it was the first event for a patient.
732   The sum of the first events across all categories will equal the total events for the
733   composite endpoint.
734   • The other approach considers the events of each type in each patient. With this method,
735   each of the components can also be treated as a distinct endpoint, irrespective of the order
736   of occurrence, giving the numbers of patients who ever experienced an event of each
737   type. In this case, each patient can be included in the event counts for more than one
738   component, and the sum of events on all component types will be greater than the total
739   number of composite events using only the first events.
740

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

756

757 **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)	

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

761

762 The second analysis showing the results for any occurrence of individual components is quite
763 different from the first-event-only decomposition analysis. There are now more total events,
764 because some patients experience more than one event type and these patients are included in
765 both component-event counts. In this example, ESRD events at any time yield a hazard ratio of
766 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.

767   0.93. Thus, the decomposition analysis limited to first events does not fully characterize the
768   effect of losartan on ESRD.
769
770   The analysis of any occurrence of an event type, however, can be complicated by the issue
771   known broadly in statistics as competing risks. This is the phenomenon wherein occurrence of
772   certain endpoints can make it impossible to observe other events in the same patient. For
773   example, in the RENAAL trial, patients whose first event was death could never be observed to
774   have doubling of serum creatinine. If one study group had higher early mortality, it could appear
775   to have a favorable profile with respect to other endpoint events simply because fewer patients
776   survived, diminishing the number of patients at risk for the other types of events.
777
778   Study design and patient management issues can also complicate interpretation of the
779   decomposition analyses. For example, in some trials, experiencing any endpoint event is cause
780   to remove a patient from study therapy and to initiate treatment with alternative agents, including
781   the possibility of receiving another treatment in the trial. Such a change in therapy obscures the
782   relationship between the initial study therapy and the occurrence of subsequent events, so that
783   only the analysis of first event will be useful. The complexities of interpretation of the
784   decomposition analyses are important to consider when planning studies with a composite
785   endpoint.
786
787   2.   *Reporting and Interpreting the Individual Component Results of a Composite*
788   *Endpoint*
789
790   The different components of a composite endpoint are selected because they are all clinically
791   important; however, because each one is not necessarily equally affected by the drug, it is
792   relevant and important to examine the effects of the drug on the individual components as well as
793   on the overall endpoint. Presenting only data on the composite might imply meaningful
794   treatment effects on all of the individual components, when a composite effect may in fact be
795   established with little or no evidence of effect on some of the individual components. On the
796   other hand, showing the results of the analysis for each of the individual components may imply
797   an effect on an individual component when an appropriate statistical analysis would not support
798   that conclusion. Thus, it is important to present descriptive analyses of between-group
799   differences for the components in a way that does not overstate the conclusions.
800
801   It is common for one component of a composite endpoint to overly influence the treatment
802   effect, but even if that is not so, and all components contribute, the inclusion of a particular
803   component in a composite does not usually support an independent conclusion of efficacy on that
804   component. FDA's guidance for industry *Clinical Studies Section of Labeling for Human*
805   *Prescription Drug and Biological Products — Content and Format*[12] calls for presentation in
806   labeling of the components of a composite endpoint but without a statistical analysis of the
807   separate components unless the components were prespecified as separate endpoints and
808   assessed with a prospectively defined hypothesis and statistical analysis plan. In such a case, the
809   statistical analysis will usually consider all events of each type, not just first-occurring events (as
810   illustrated in Table 1 above). Only findings on prespecified endpoints that are statistically

---

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

811 significant, with adjustment for multiplicity, are considered demonstrated effects of a drug. All
812 other findings are considered descriptive and would require further study to demonstrate that
813 they are true effects of the drug. For example, a composite endpoint that includes mortality as a
814 component provides little information about effects on mortality if there are few deaths, and
815 presentations can make that clear by showing the actual numbers of deaths. Therefore, clear
816 presentation of the results of the components of a composite is essential to describe where the
817 drug's effect occurs. For example, the LIFE trial comparing losartan and atenolol in people with
818 hypertension showed a clear, statistically-significant advantage of losartan on the composite
819 endpoint of death, nonfatal myocardial infarction, or stroke, but this appeared to be related to an
820 effect on fatal and nonfatal stroke, with no advantage on the incidence of acute myocardial
821 infarction or cardiovascular death.[13]
822
823 To demonstrate an effect on a specific component or components of a composite endpoint, the
824 component or components should be included prospectively as a secondary endpoint for the
825 study or possibly as an additional primary endpoint (see section III.C.5), with appropriate Type I
826 error rate control. If control of the Type I error rate is ensured with respect to the individual
827 component or components, in addition to control for the composite, a trial will be potentially
828 able to support conclusions regarding drug effects on the individual component or components as
829 well as the composite.
830
831   *3.   Evaluating and Reporting the Results on Other Multi-Component Endpoints*
832
833 As with composite endpoints, understanding which components of a within-patient multi-
834 component endpoint (e.g., symptom rating scale such as HAM-D) have contributed most to the
835 overall statistical significance could be important to correctly understanding the clinical effects
836 of the drug. Consequently, a descriptive analysis of the study results on the individual
837 components (or, in some cases, groups of similar components) may be considered but, as stated
838 previously, if undertaken, should be presented in a way that does not overstate the conclusions.
839 Unlike the composite endpoint used for outcome studies, where each component usually has
840 clear clinical importance (death, acute myocardial infarction, stroke, hospitalization), the clinical
841 importance of the components of these patient assessments may be less clear. Thus, for many of
842 these multi-component endpoints, the overall score is regarded as comprehensive and clinically
843 interpretable. The individual components of the scales, however, may not be independently
844 clinically interpretable. Although some rating scales have been developed with broad
845 multicomponent domains to allow the domains to be interpretable subsets of the overall scale,
846 the individual domain and subscale scores generally are not prespecified for hypothesis testing.
847 Prespecification of subscale scores with appropriate multiplicity control is required if it is
848 thought to be important to demonstrate an effect of a drug on one or more of these subscale
849 scores in addition to the overall multi-component endpoint.
850
851 Analyses of specific component item(s) of a symptom rating scale as explicit endpoints in the
852 primary or secondary endpoint families may be reasonable, contingent on being clinically
853 interpretable, in two cases:

---

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

854   (1) where earlier trials have suggested targeted efficacy of a drug on one or a small
855   number of specific symptoms, or
856   (2) where the specific symptom measured by the item is considered to be of substantial
857   inherent clinical importance.
858
859 An example of the first type is a novel agent for rheumatoid arthritis that was found in a
860 controlled phase 2 trial to be particularly effective in lessening patients' pain. In this example, a
861 sponsor might wish to test this hypothesis using a pain scale as a secondary endpoint in a trial
862 where improvement meeting ACR20 criteria, which include pain as a component, is the primary
863 endpoint. An example of the second type of component analysis might be found in trials of anti-
864 psychotic drugs, in which positive and negative symptoms are domains collected in the Positive
865 and Negative Syndrome Scale (PANSS) and often analyzed separately in addition to the overall
866 scale. Interpretation of analyses of any subscale domain, however, is dependent on that subscale
867 domain having been previously evaluated and determined to be valid as a stand-alone clinical
868 measure. As described above (see section III.C), control of the Type I error rate will still be
869 necessary for both the primary and secondary endpoint families.
870
871
872 **IV.   STATISTICAL METHODS**
873
874 A variety of situations in which multiplicity arises have been discussed in sections II and III.
875 Statistical methods provide acceptable ways to correct for multiplicity and control the Type I
876 error rate for many of them. Standard statistical methods are available, for example:
877   • to examine treatment effects for multiple endpoints where success on any one endpoint
878   would be acceptable, and
879   • to allow sequential testing where success on one endpoint permits analysis of additional
880   endpoints.
881
882 This section describes methods that are commonly used for handling multiplicity problems in
883 controlled clinical trials that examine treatment effects on multiple endpoints.
884
885   **A.   Type I Error Rate for a Family of Endpoints and Conclusions on Individual**
886   **Endpoints**
887
888 When there is a family of endpoints (discussed in sections II.A and III.A), the Type I error rate
889 commonly used for the group of study endpoints is called the family-wise Type I error rate
890 (FWER) or the overall Type I error rate for the family. The FWER is the probability of
891 erroneously finding a statistically-significant treatment effect in at least one endpoint regardless
892 of the presence or absence of treatment effects in the other endpoints within the family. This
893 error rate is typically held to 0.05 (0.025 for one-sided tests). The statistical methods discussed
894 in section IV.C maintain control of the FWER for finding significant treatment effects for study
895 endpoints individually, thereby permitting an individual effectiveness conclusion on each
896 endpoint.
897
898 There are also other statistical analysis methods, often called global procedures, that control the
899 FWER with regard to erroneously concluding that there is a treatment effect on some endpoint

900 (one or more) when there is no such effect on any endpoint. These methods allow a conclusion
901 of treatment effectiveness in the global sense, but do not support reaching conclusions on the
902 individual endpoints within the family. These methods are generally not encouraged when study
903 designs and methods that test the endpoints individually are feasible; therefore, these global
904 procedures are not described in this guidance.
905
906 Because composite and other multi-component endpoints (see sections III.C.3 and III.C.4) are
907 constructed as a single endpoint, when they are part of an endpoint family, the methods
908 described in section IV.C can be applied to them.
909

910 **B.   When the Type I Error Rate Is Not Inflated or When the Multiplicity**
911 **Problem Is Addressed Without Statistical Adjustment or by Other Methods**
912
913 This section identifies two situations involving multiple endpoints where inflation of the Type I
914 error rate is avoided so that adjustments for multiplicity are not needed. These situations assume
915 that the trial has no interim analysis or mid-course design modifications.
916

917 *1.   Clinically Relevant Benefits Required for All Specified Primary Endpoints — the*
918 *Case of "Co-Primary" Endpoints*[14]
919
920 As discussed in detail in section III.C, when multiple primary endpoints are tested and success in
921 the study depends on success on all endpoints (i.e., they are co-primary endpoints), no
922 multiplicity adjustment is necessary because there is no opportunity to select the most favorable
923 result from among several endpoints. The impact of multiplicity in these situations is to increase
924 the Type II error rate (section III.B).
925

926 *2.   Use of Multiple Analyses Methods for a Single Endpoint after Success on the*
927 *Prespecified Primary Analysis Method*
928
929 For many trials there are a range of plausible, closely related analyses of an individual endpoint.
930 For example, the primary analysis of an outcome trial could adjust for certain covariates, make a
931 different choice of covariates, make no covariate adjustment, be conducted on the intent-to-treat
932 (ITT) population or various modified populations, or use various hypothesis testing methods.
933 Accepting any one of these multiple analyses, when successful, as a basis for a conclusion that
934 there is a treatment effect would increase the study Type I error rate, but it is difficult to estimate
935 the increase in error rate because the results of these different analyses are likely to be similar
936 and it is unclear how many choices could have been made. As with other multiplicity problems,
937 prospective specification of the analysis method will generally eliminate the concern about a
938 biased (result-driven) choice.
939
940 Once the effect has been clearly demonstrated based on the prespecified primary analysis,
941 alternative analyses of the primary endpoint may be needed to correctly interpret the study's
942 results. Additional analyses of the primary endpoint may be needed to gain a better

---

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

943 understanding of the observed treatment effect (e.g., to use a less conservative analysis to better
944 estimate the effect size). In other cases, multiple related analyses are used to assess the
945 sensitivity of the results to the important underlying assumptions of the prespecified analysis
946 method. For example, sensitivity analyses may be needed to determine the impact of missing
947 data on the primary analysis results, when the primary analysis method relies on unverifiable
948 assumptions about those missing data. Note that these additional analyses do not demonstrate
949 any new effects of the drug; rather, they clarify the effect already demonstrated by the primary
950 analysis of a successful study.
951
952 **C.   Common Statistical Methods for Addressing Multiple Endpoint-Related**
953 **Multiplicity Problems**
954
955 This section presents some common statistical methods and approaches for addressing
956 multiplicity problems in controlled clinical trials that evaluate treatment effects on multiple
957 endpoints. The choice of the method to use for a specific clinical trial will depend on the
958 objectives and the design of the trial, as well as the knowledge of the drug being developed and
959 the clinical disorder. The method, however, should be decided upon prospectively. Because the
960 considerations that go into the choice of multiplicity adjustment method can be complex and
961 specific to individual product development programs, this guidance does not attempt to
962 recommend any one method over another in most cases. Sponsors should consider the variety of
963 methods available and in the prospective analysis plan select the most powerful method that is
964 suitable for the design and objective of the study and maintains Type I error rate control. There
965 are, for example, a small number of situations in which one method is unambiguously more
966 powerful than another without inflating the Type I error rate beyond the nominal level (e.g., the
967 Holm method is more powerful than the Bonferroni method for primary endpoints). These
968 situations are noted below.
969
970 The methods presented here are general, and the discussions and hypothetical examples have
971 been generally limited to two-arm trials that examine treatment versus control differences on
972 multiple endpoints. Similar considerations may apply to other kinds of multiplicity, such as in
973 assessing treatment effects at different time points, or at different doses. Although the following
974 discussions are oriented to the general reader, application of many of these methods can be
975 technically complex and should be used relying on statistical expertise. Consequently, when a
976 multiple endpoints problem arises in designing a clinical trial and one or more of these methods
977 are to be considered, consultation with knowledgeable experts is important.
978
979 Statistical methods for addressing multiplicity issues are broadly classified into two types:
980 single-step and multistep procedures. Single-step procedures provide for parallel (simultaneous)
981 testing and simultaneous (adjusted) confidence intervals for assessing the magnitude of the
982 treatment effects. Single-step procedures tend to cause loss of study power, so that sample sizes
983 need to be increased in comparison to sample sizes needed for a single-endpoint study.
984 Multistep procedures are generally more efficient in that they better preserve the power of the
985 tests, but do not readily provide adjusted confidence intervals. There are several kinds of
986 multistep procedures, for example step-down, step-up, and sequential procedures.
987

988   In a step-down procedure, one calculates the p-values from all tests to be considered at one time
989   and starts hypothesis testing with the smallest p-value (i.e., statistically the most robust endpoint
990   test) and then steps down to the next smallest p-value (i.e., the next most robust endpoint test),
991   and so on. In a step-up procedure, one proceeds in the reverse direction. That is, one starts with
992   the largest p-value (i.e., the least robust test) and steps up to the second-largest p-value, finally
993   reaching the smallest p-value (i.e., the most robust test). These approaches are covered in the
994   following sections; e.g., the Holm procedure is a step-down procedure and the Hochberg
995   procedure is a step-up procedure.
996
997   *1.*   *The Bonferroni Method*
998
999   The Bonferroni method is a single-step procedure that is commonly used, perhaps because of its
1000   simplicity and broad applicability. It is a conservative test and a finding that survives a
1001   Bonferroni adjustment is a credible trial outcome. The drug is considered to have shown effects
1002   for each endpoint that succeeds on this test. The Holm (section IV.C.2) and Hochberg (section
1003   IV.C.3) methods are more powerful than the Bonferroni method for primary endpoints and are
1004   therefore preferable in many cases. However, for reasons detailed in sections IV.C.2-3, sponsors
1005   may still wish to use the Bonferroni method for primary endpoints in order to maximize power
1006   for secondary endpoints or because the assumptions of the Hochberg method are not justified.
1007
1008   The most common form of the Bonferroni method divides the available total alpha (typically
1009   0.05) equally among the chosen endpoints. The method then concludes that a treatment effect is
1010   significant at the alpha level for each one of the *m* endpoints for which the endpoint's p-value is
1011   less than α/m. Thus, with two endpoints, the critical alpha for each endpoint is 0.025, with four
1012   endpoints it is 0.0125, and so on. Therefore, if a trial with four endpoints produces two-sided p-
1013   values of 0.012, 0.026, 0.016, and 0.055 for its four primary endpoints, the Bonferroni method
1014   would compare each of these p-values to the divided alpha of 0.0125. The method would
1015   conclude that there was a significant treatment effect at level 0.05 for only the first endpoint,
1016   because only the first endpoint has a p-value of less than 0.0125 (0.012). If two of the p-values
1017   were below 0.0125, then the drug would be considered to have demonstrated effectiveness on
1018   both of the specific health effects evaluated by the two endpoints.
1019
1020   The Bonferroni method tends to be conservative for the study overall Type I error rate if the
1021   endpoints are positively correlated, especially when there are a large number of positively-
1022   correlated endpoints. Consider a case in which all of three endpoints give nominal p-values
1023   between 0.025 and 0.05, i.e., all 'significant' at the 0.05 level but none significant under the
1024   Bonferroni method. Such an outcome seems intuitively to show effectiveness on all three
1025   endpoints, but each would fail the Bonferroni test. When there are more than two endpoints
1026   with, for example, correlation of 0.6 to 0.8 between them, the true family-wise Type I error rate
1027   may decrease from 0.05 to approximately 0.04 to 0.03, respectively, with negative impact on the
1028   Type II error rate. Because it is difficult to know the true correlation structure among different
1029   endpoints (not simply the observed correlations within the dataset of the particular study), it is
1030   generally not possible to statistically adjust (relax) the Type I error rate for such correlations.
1031   When a multiple-arm study design is used (e.g., with several dose-level groups), there are
1032   methods that take into account the correlation arising from comparing each treatment group to a
1033   common control group.

1034
1035   The Bonferroni test can also be performed with different weights assigned to endpoints, with the
1036   sum of the relative weights equal to 1.0 (e.g., 0.4, 0.1, 0.3, and 0.2, for four endpoints). These
1037   weights are prespecified in the design of the trial, taking into consideration the clinical
1038   importance of the endpoints, the likelihood of success, or other factors. There are two ways to
1039   perform the weighted Bonferroni test:
1040
1041   •   The unequally weighted Bonferroni method is often applied by dividing the overall alpha
1042   (e.g., 0.05) into unequal portions, prospectively assigning a specific amount of alpha to
1043   each endpoint by multiplying the overall alpha by the assigned weight factor. The sum of
1044   the endpoint-specific alphas will always be the overall alpha, and each endpoint's
1045   calculated p-value is compared to the assigned endpoint-specific alpha.
1046
1047   •   An alternative approach is to adjust the raw calculated p-value for each endpoint by the
1048   fractional weight assigned to it (i.e., divide each raw p-value by the endpoint's weight
1049   factor), and then compare the adjusted p-values to the overall alpha of 0.05.
1050
1051   These two approaches are equivalent.
1052
1053   *2.   The Holm Procedure*
1054
1055   The Holm procedure is a multi-step step-down procedure; it is useful for endpoints with any
1056   degree of correlation. It is less conservative than the Bonferroni method because a success with
1057   the smallest p-value (at the same endpoint-specific alpha as the Bonferroni method) allows other
1058   endpoints to be tested at larger endpoint-specific alpha levels than does the Bonferroni method.
1059   The algorithm for performing this test is as follows:
1060
1061   The endpoint p-values resulting from the completed study are first ordered from the smallest to
1062   the largest. Suppose that there are m endpoints to be tested and $p_{(1)}$ represents the smallest p-
1063   value, $p_{(2)}$ the next-smallest p-value, $p_{(3)}$ the third-smallest p-value, and so on.
1064
1065   i.   The test begins by comparing the smallest p-value, $p_{(1)}$, to α/m, the same threshold used
1066   in the equally-weighted Bonferroni correction. If this $p_{(1)}$ is less than α/m, the treatment
1067   effect for the endpoint associated with this p-value is considered significant.
1068
1069   ii. The test then compares the next-smallest p-value, $p_{(2)}$, to an endpoint-specific alpha of
1070   the total alpha divided by the number of yet-untested endpoints (e.g., $\alpha/[m-1]$ for the
1071   second smallest p-value, a somewhat less conservative significance level). If $p_{(2)} <$
1072 $\alpha/(m-1)$, then the treatment effect for the endpoint associated with this $p_{(2)}$ is also
1073   considered significant.
1074
1075   iii. The test then compares the next ordered p-value, $p_{(3)}$, to $\alpha/(m-2)$, and so on until the last
1076   p-value (the largest p-value) is compared to α.
1077

1078   iv.   The procedure stops, however, whenever a step yields a non-significant result. Once an
1079   ordered p-value is not significant, the remaining larger p-values are not evaluated and it
1080   cannot be concluded that a treatment effect is shown for those remaining endpoints.
1081
1082   For example, when $\alpha = 0.05$, and there are four endpoints (m = 4), the significance level for the
1083   smallest p-value is $\alpha/m = 0.05/4 = 0.0125$, and significance levels for the subsequent ordered p-
1084   values are $\alpha/(m-1) = 0.05/3 = 0.0167$, $\alpha/(m-2) = 0.05/2 = 0.025$, and $\alpha/(m-3) = 0.05/1 = 0.05$,
1085   respectively.
1086
1087   To illustrate, we apply the Holm procedure to the two-sided study result p-values used to explain
1088   the Bonferroni method: 0.012, 0.026, 0.016, and 0.055 associated with endpoints one to four,
1089   respectively (p_1, p_2, p_3, p_4). With four endpoints, the successive endpoint-specific alphas are
1090   0.0125, 0.0167, 0.025, and 0.05. The smallest p-value in this group is $p_1 = 0.012$, which is less
1091   than 0.0125. The treatment effect for endpoint one is thus successfully demonstrated and the test
1092   continues to the second step. In the second step, the second smallest p-value is $p_3 = 0.016$, which
1093   is compared to 0.0167. Endpoint three has therefore also successfully demonstrated a treatment
1094   effect, as 0.016 is less than 0.0167. Testing is now able to proceed to the third step, in which the
1095   next ordered p-value of $p_2 = 0.026$ is compared to 0.025. In this comparison, as 0.026 is greater
1096   than 0.025, the test is not statistically significant. This non-significant result stops further tests.
1097   Therefore, in this example, this procedure concludes that treatment effects have been shown for
1098   endpoints one and three.
1099
1100   As noted, the Holm procedure is less conservative (and thereby more powerful) than the
1101   Bonferroni test. It tests the smallest p-value at the same alpha as the Bonferroni test, but, given a
1102   statistically-significant result on that endpoint, it tests subsequent p-values at higher significance
1103   levels. In the above example, the Bonferroni test was able to conclude that there is a significant
1104   treatment effect at the overall level 0.05 for endpoint one only; the Holm test was able to do so
1105   for endpoints one and three. Both, however, require at least one endpoint with a p-value <
1106 $0.05/m$. The Holm procedure is also more flexible than simple prospective ordering of endpoints
1107   for testing (section IV.C.5). It allows testing of the endpoint with the smallest p-value first,
1108   without knowing in advance which endpoint that will be. A disadvantage of the Holm procedure
1109   is the potential inability to pass along *unused alpha* (see section IV.C.6) to a secondary endpoint
1110   family because testing of any additional endpoints is not permitted when one of the sequentially-
1111   tested endpoints in the family fails to reject the null hypothesis.
1112
1113   *3.   The Hochberg Procedure*
1114
1115   The Hochberg procedure is a multi-step, step-up testing procedure. It compares the p-values to
1116   the same alpha critical values of α/m, $\alpha/(m-1)$, …, α, as the Holm procedure, but, in contrast to
1117   the Holm procedure, the Hochberg procedure is a step-up procedure. Instead of starting with the
1118   smallest p-value, the procedure starts with the largest p-value, which is compared to the largest
1119   endpoint-specific critical value (α). Also, essentially in the reverse of the Holm procedure, if the
1120   first test of hypothesis does not show statistical significance, testing proceeds to compare the
1121   second-largest p-value to the second-largest adjusted alpha value, $\alpha/2$. Sequential testing
1122   continues in this manner until a p-value for an endpoint is statistically significant, whereupon the
1123   Hochberg procedure provides a conclusion of statistically-significant treatment effects for that

1124 endpoint and all endpoints with smaller p-values. For example, when the largest p-value is less
1125 than α, then the method concludes that there are significant treatment effects for all endpoints. In
1126 another situation, when the largest p-value is not less than α, but the second-largest p-value is
1127 less than $\alpha/2$, then the method concludes that treatment effects have been demonstrated for all
1128 endpoints except for the one associated with the largest p-value.
1129
1130 To illustrate, consider the same two-sided p-values used in the previous examples: 0.012, 0.026,
1131 0.016, and 0.055 associated with endpoints one to four, respectively (p_1, p_2, p_3, p_4).
1132

1133   i.   The largest p-value of $p_4 = 0.055$ is compared to its alpha critical value of $\alpha = 0.05$.
1134   Because this p-value of 0.055 is greater than 0.05, the treatment effect for the endpoint
1135   four associated with this p-value is considered not significant. The procedure,
1136   however, continues to the second step.

1137   ii. In the second step, the second largest p-value, $p_2 = 0.026$, is compared to $\alpha/2 = 0.025$;
1138 p_2 is also greater than the allocated alpha, and endpoint two associated with this p-value
1139   is also not statistically significant.

1140   iii. In the third step, the next largest p-value, $p_3 = 0.016$, is compared to its alpha critical
1141   value of $\alpha/3 = 0.0167$, and this endpoint does show a significant treatment effect.

1142   iv. The significant result on endpoint three automatically causes the treatment effect for all
1143   untested endpoints (which will have smaller p-values) to be significant as well (i.e.,
1144   endpoint one in this case).
1145
1146 Although for this specific example, the endpoints that are statistically significant are the same as
1147 for the Holm procedure, the Hochberg procedure is potentially more powerful. The Hochberg
1148 procedure may conclude that there are significant treatment effects for more endpoints than
1149 would the Holm procedure, depending on the specific p-values obtained in the study. This is
1150 because the Hochberg procedure allows testing of endpoints from the largest p-value to the
1151 smallest and concludes that all remaining endpoints are successful as soon as one test is
1152 successful, even if the remaining p-values would not have succeeded on testing with their
1153 appropriate sequential alpha level. In contrast, the Holm procedure tests from smallest p-value to
1154 largest and determines that all untested endpoints are unsuccessful as soon as one test is
1155 unsuccessful, even if those remaining endpoints would have been successful if tested with their
1156 appropriate sequential alpha level.
1157
1158 Thus, for the case of two endpoints, if the two-sided p-values were 0.026 and 0.045, the
1159 Hochberg procedure will conclude that there are significant treatment effects on both endpoints,
1160 but the Holm procedure will fail on both. In the Hochberg procedure, the larger of the two p-
1161 values, $p = 0.045$ ($< \alpha = 0.05$), is a significant result, and the second endpoint is automatically
1162 considered significant. In the Holm procedure, the smaller of the two p-values, 0.026 ($> \alpha/m =$
1163 0.05/2), is a non-significant result; therefore, the larger p-value is not evaluated.
1164
1165 The Bonferroni and the Holm procedures are well known for being assumption-free. The
1166 methods can be applied without concern for the endpoint types, their statistical distributions, and
1167 the type of correlation structure. The Hochberg procedure, on the other hand, is not assumption-

1168   free in this way. The Hochberg procedure is known to provide adequate overall alpha-control for
1169   independent endpoint tests and also for two positively-correlated dependent tests with standard
1170   test statistics, such as the normal Z, student's t, and 1 degree of freedom chi-square. It is also a
1171   valid test procedure when certain conditions are met. Various simulation experiments for the
1172   general case (e.g., for more than two endpoints with unequal correlation structures) indicate that
1173   the Hochberg procedure usually will, but is not guaranteed to, control the overall Type I error
1174   rate for positively-correlated endpoints, but fails to do so for some negatively-correlated
1175   endpoints. Therefore, beyond the aforementioned cases where the Hochberg procedure is known
1176   to be valid, its use is generally not recommended for the primary comparisons of confirmatory
1177   clinical trials unless it can be shown that adequate control of Type I error rate is provided.
1178
1179   *4.*   *Prospective Alpha Allocation Scheme*
1180
1181   The Prospective Alpha Allocation Scheme (PAAS) is a single-step method that has a slight
1182   advantage in power over the Bonferroni method. The method allows equal or unequal alpha
1183   allocations to all endpoints, but, as with the Bonferroni method, each specific endpoint must
1184   receive a prospective allocation of a specific amount of the overall alpha. The alpha allocations
1185   are required to satisfy the equation:
1186
1187 $(1 - \alpha_1)(1 - \alpha_2) \dots (1 - \alpha_k) \dots (1 - \alpha_m) = (1 - \alpha)$.
1188
1189   Each element in this equation, $(1 - \alpha_k)$, is the probability of correctly not rejecting the null
1190   hypothesis for the kth endpoint, when it is tested at the allocated alpha α_k. When the Type I error
1191   rate for the study is set at 0.05 overall, the probability of correctly not rejecting any of the
1192   individual null hypotheses (i.e., when all null hypotheses are true) must be 1- 0.05 = 0.95 = (1-
1193   alpha). This equation states the requirement that probability of correctly not rejecting all of the
1194   individual null hypotheses, calculated by multiplying each of the m probabilities together, must
1195   equal the selected goal (e.g., 0.95). The alpha allocation for any of the individual endpoint tests
1196   can be arbitrarily assigned, if desired, but the total group of allocations must always satisfy the
1197   above equation. In general, when arbitrary alpha allocations are made for some endpoints, at
1198   least the last endpoint's alpha must be calculated in order to satisfy the overall equation. As
1199   stated earlier, the Bonferroni method relies upon a similar constraint-defining equation, except
1200   that for the Bonferroni method the sum of all the individual alphas must equal the overall study-
1201   wise alpha.
1202
1203   Consider the case of three endpoints with two arbitrary alpha allocations in which $\alpha_1 = 0.02$ and
1204 $\alpha_2 = 0.025$ are assigned to the first two endpoints. If the total $\alpha = 0.05$, then the third endpoint
1205   would have an alpha of 0.0057, because the above equation becomes $(0.98)(0.975)(1 - \alpha_3) =$
1206   0.95, so that $\alpha_3 = 0.0057$ for the third endpoint, instead of 0.005, as would have been assigned by
1207   the Bonferroni method $(0.02 + 0.025 + 0.005 = 0.05)$. When all alpha allocations are equal, then
1208   the individual comparison alpha is given by $1 - (1 - \alpha)^{1/m}$. This adjustment formula is also known
1209   as the Šidák adjustment formula. For the case of three endpoints, this adjusted alpha is 0.01695,
1210   which is only slightly greater than the 0.0167 assigned by the Bonferroni method. The slight
1211   savings in alpha provides a slight gain in the power of the tests. The PAAS ensures FWER
1212   control for all comparisons that are independent or positively correlated. If the endpoints are
1213   negatively correlated, FWER control may not be assured.

1214
1215   5.   *The Fixed-Sequence Method*
1216
1217   The multiplicity problem arises from conducting tests for each of the multiple endpoints where
1218   each test provides an opportunity to decide that the study was successful. Any method that
1219   adequately adjusts for the multiplicity of opportunities will address the problem. In many
1220   studies, testing of the endpoints can be ordered in a specified sequence, often ranking them by
1221   clinical relevance or likelihood of success. A fixed-sequence statistical strategy tests endpoints
1222   in a predefined order, all at the same significance level alpha (e.g., $\alpha = 0.05$), moving to a second
1223   endpoint only after a success on the previous endpoint. Such a test procedure does not inflate the
1224   Type I error rate as long as there is (1) prospective specification of the testing sequence and (2)
1225   no further testing once the sequence breaks, that is, further testing stops as soon as there is a
1226   failure of an endpoint in the sequence to show significance at level alpha (e.g., $\alpha = 0.05$).
1227
1228   The idea behind this sequential testing method is that when there is a significant treatment effect
1229   for an endpoint, then the alpha level for this test remains available to be carried forward (passed
1230   along) to the next endpoint test in the sequence. However, the method uses all of the available
1231   alpha as soon as a non-significant result occurs. The order of testing is therefore critical.
1232
1233   The statistical conclusions provided by this method may differ from those provided by other
1234   methods, and they depend on the ordering of the tests. Consider, for example, a trial with three
1235   primary endpoints, A, B, and C, whose two-sided *p*-values for treatment effects are: $p_A = 0.045$,
1236 $p_B = 0.016$ and $p_C = 0.065$. This trial would conclude that there was a significant treatment effect
1237   for only the endpoint B by the Bonferroni test, because $p_B = 0.016 < 0.0167$ (i.e., 0.05/3), but
1238   would not conclude that there was a significant effect on endpoints A or C. The Holm test would
1239   not find significant effects for additional endpoints either, unless the p-value for endpoint A was
1240 $p < 0.025$. If the study had planned sequential testing in the order of (C, B, A), it would be an
1241   entirely failed study, because $p_C = 0.065 > 0.05$, and no further testing would be performed after
1242   the first failed test for endpoint C. On the other hand, this trial would show significant treatment
1243   effects for endpoints B and A if it had planned sequential testing in the order of (B, A, C),
1244   because $p_B = 0.016 < 0.05$, and following it, $p_A = 0.045 < 0.05$; the same effects would be shown
1245   if the order was (A, B, C). Thus, the fixed-order sequential testing method has the potential to
1246   find more endpoints successful than the single-step methods, but it also has the potential to find
1247   fewer endpoints successful, depending on the order chosen.
1248
1249   The appeal of the fixed-sequence testing method is that it does not require any alpha adjustment
1250   of the individual tests. Its main drawback is that if a hypothesis in the sequence is not rejected, a
1251   statistical conclusion cannot be made about the endpoints planned for the subsequent hypotheses,
1252   even if they have extremely small p-values. Suppose, for example, that in a study, the p-value
1253   for the first endpoint test in the sequence is $p = 0.250$, and the p-value for the second endpoint is
1254 $p = 0.0001$; despite the apparent "strong" finding for the second endpoint, no formal favorable
1255   statistical conclusion can be reached for this endpoint. Although it may seem counterintuitive to
1256   ignore such an apparently strong result, to allow a conclusion of drug effectiveness based on the
1257   second endpoint would in fact be ignoring the first endpoint's result and returning to the situation
1258   of having multiple separate opportunities to declare the study a success. Such a post hoc rescue

1259 recreates the multiplicity problem, and causes inflation of the study-wise Type I error rate. The
1260 example discussed here would, of course, have shown an effect using a Bonferroni test.
1261
1262 Thus, carefully selecting the ordering of the tests of hypotheses is essential. A test early in the
1263 sequence that fails to show statistical significance will render the remainder of the endpoints not
1264 statistically significant. It is often not possible to determine a priori the best order for testing,
1265 and there are other methods for addressing the multiplicity problem, which are described in the
1266 following subsections.
1267
1268   *6.   The Fallback Method*
1269
1270 The fallback method is a modification of the fixed-sequence method that provides some
1271 opportunity to test an endpoint later in the sequence even if an endpoint tested early in the
1272 sequence has failed to show statistical significance. The order of the endpoints remains
1273 important. The appeal of the fallback method is that if an endpoint later in the sequence has a
1274 robust treatment effect while the preceding endpoint is unsuccessful, there is a modest amount of
1275 alpha retained as a fallback to allow interpretation of that endpoint without inflating the Type I
1276 error rate.
1277
1278 Applying the fallback method begins by dividing the total alpha (not necessarily equally) among
1279 the endpoints, and maintains a fixed sequence for the testing. As the testing sequence
1280 progresses, a successful test preserves its assigned alpha as "saved" (unused) alpha that is passed
1281 along to the next test in the sequence, as is the case for the sequential method. This passed-along
1282 alpha is added to the prospectively assigned alpha (if any) of that next endpoint and the summed
1283 alpha is used for testing that endpoint. Thus, as sequential tests are successful, the alpha
1284 accumulates for the endpoints later in the sequence; these endpoints are then tested with
1285 progressively larger alphas.
1286
1287 To illustrate, consider a cardiovascular trial in which the first primary endpoint is exercise
1288 capacity, for which the trial is adequately powered. The second primary endpoint is mortality,
1289 for which the trial is underpowered.
1290
1291   i. Under the fallback method, we may assign $\alpha_1 = 0.04$ for the first endpoint test and save
1292   alpha of 0.01 for the second endpoint test. Any other desired division of the available
1293   overall alpha would also be permitted.
1294
1295   ii. If the first endpoint test is significant at level $\alpha_1 = 0.04$, this alpha is unused and is
1296   passed to the second endpoint test as an additional alpha of 0.04, giving a total alpha for
1296   the second endpoint test of 0.05 (0.01 + 0.04). The second endpoint test is then
1297   performed at the significance level of 0.05.
1298
1298   iii. If the first endpoint is not significant at level 0.04, then this alpha of 0.04 is not
1299   available to be passed on for the second endpoint test. The test for the second endpoint
1300   is at the originally reserved alpha of 0.01.
1301
1302 In practice, users of this method usually assign most of the alpha to the first primary endpoint
1303 and the remainder to the second endpoint, although other distributions are also valid. The

1304 fallback method is often used when there is an endpoint thought less likely than another to be
1305 statistically significant, so that it is not designated the first endpoint, but is nevertheless of
1306 substantial clinical importance. The fallback method could conclude that an unexpectedly robust
1307 finding is statistically interpretable as a positive result even if the first primary endpoint failed,
1308 without inflation of the Type I error rate.
1309
1310 The statistical power of the fallback method depends primarily on the magnitude of the effect on,
1311 and alpha assigned to, each of the ordered endpoints. As with the simple fixed-sequence method,
1312 the overall power of the fallback method exceeds that of the Bonferroni test, because when the
1313 earlier endpoints show significant results, the method uses larger alpha levels for later endpoints
1314 than is possible under the Bonferroni method.
1315
1316   *7.   Gatekeeping Testing Strategies*
1317
1318 Clinical trials commonly assess efficacy of a treatment on multiple endpoints, usually grouped
1319 into a primary endpoint or endpoint family, and a secondary endpoint or endpoint family (see
1320 sections II.A and III.A). The usual strategy is to test all endpoints in the primary family
1321 according to one of the previously discussed methods (e.g., Bonferroni, fallback) and proceed to
1322 the secondary family of endpoints only if there has been statistical success in the primary family.
1323 This allows all of the available alpha level to be distributed within the primary family
1324 (containing the most important study endpoints) and thus maximizes the study power for those
1325 endpoints. In contrast, if the available alpha were distributed among all of the endpoints in the
1326 primary and secondary families, power would be reduced for the primary endpoints. Although it
1327 is not generally recommended, if there were an additional family of endpoints for which it was
1328 also important to control the Type I error rate, that family could be designated as third in the
1329 sequence.
1330
1331 This approach of testing the primary family first, and then the secondary family contingent upon
1332 the results within the primary family is called the gatekeeping testing strategy to highlight the
1333 fact that the endpoint families are analyzed in a sequence, with each family serving as a
1334 gatekeeper for the next one. The tests for the secondary family (and subsequent families if any)
1335 are carried out with appropriate multiplicity adjustments within that family, but only if the tests
1336 in the primary family have been successful.
1337
1338 Two types of gatekeeping testing strategies are common in clinical trials, serial and parallel,
1339 determined by how the endpoints are tested within the primary family. The term serial strategy
1340 is applied when the endpoints of the primary family are tested as co-primary endpoints (section
1341 III.C). If all endpoints in the primary family are statistically significant at the same alpha level
1342 (e.g., $\alpha = 0.05$), the endpoints in the second family are examined. The endpoints in the second
1343 family are tested by any one of several possible methods (e.g., Holm procedure, the fixed-
1344 sequence method, or others described in section IV.C). If, however, at least one of the null
1345 hypotheses of the primary family fails to be rejected, the primary family criterion has not been
1346 met and the secondary endpoint family is not tested.
1347
1348 The term parallel gatekeeping strategy is applied when the endpoints in the primary family are
1349 not all co-primary endpoints, and a testing method that allows the passing along of alpha from an

1350   individual test to a subsequent test (e.g., Bonferroni method or Truncated Holm method
1351   described next) is specified. In this strategy, the second endpoint family is examined when at
1352   least one of the endpoints in the first family has shown statistical significance.
1353
1354   The Bonferroni method is sometimes used for the parallel gatekeeping strategy, as it is the
1355   simplest approach. The secondary endpoint family may use a different method (e.g., the fixed-
1356   sequence method or Holm method). In this approach, if an endpoint comparison within the
1357   primary family is statistically significant at its allocated (or accumulated) endpoint-specific alpha
1358   level, then this alpha level can be validly passed on to the next family. On the other hand, if an
1359   endpoint comparison in a family is not significant at its endpoint-specific alpha level, that alpha
1360   is not passed on to the next family. The overall alpha available for testing the secondary family
1361   is the accumulated (unused) endpoint-specific alpha levels of those comparisons in the primary
1362   family that were found significant.
1363
1364   To illustrate, consider a trial whose primary objective is to test for superiority of a treatment to
1365   placebo for five endpoints: A, B, C, D and E. For this objective, the trial organizes the endpoints
1366   hierarchically into a primary family F1 = {A, B} and a secondary family F2 = {C, D, and E}.
1367   The statistical plan is to assign the total available alpha (0.05) to F1 and test the endpoints A and
1368   B in F1 by the Bonferroni method at endpoint-specific alpha levels of 0.04 and 0.01,
1369   respectively. No alpha is reserved for the second family, and the second family is tested with the
1370   Holm procedure with whatever amount of alpha is passed along to it. If, at the completion of the
1371   tests for F1, the p-values for the endpoints A and B are 0.035 and 0.055, respectively, and the p-
1372   values for endpoints C, D and E are 0.011, 0.045, and 0.019, respectively, then:
1373
1374   i.   The result for endpoint A is significant, but the result for endpoint B is not, leaving
1375   alpha of 0.04 as unused and alpha of 0.01 as used.

1376   ii. The total alpha available for testing the endpoints in F2 is 0.04 and not 0.05.

1377   iii. The endpoints C and E are significant at level 0.04 by the Holm test (C, E, and D are
1378   tested at levels of 0.0133, 0.02, 0.04, respectively).
1379
1380   The gatekeeping method described above controls the study-wise Type I error rate (e.g., at level
1381   0.05) associated with the trial's primary and secondary families. The study-wise Type I error
1382   rate takes into consideration the potential for an erroneous conclusion of efficacy for any
1383   endpoint in any family and the multiple possibilities of the drug being truly effective or
1384   ineffective on any of the endpoints. The gatekeeping strategy controls the study-wise Type I
1385   error rate when the principle of passing along only unused alpha from statistically-significant
1386   tests of hypotheses is applied. In contrast, however, independent error rate control of each
1387   family's FWER (i.e., testing each family at a separate 0.05) can lead to inflation of the study-
1388   wise Type I error rate when some, but not all, of the null hypotheses for the primary endpoint
1389   family are in fact true.
1390
1391   8.   *The Truncated Holm and Hochberg Procedures for Parallel Gatekeeping*
1392
1393   When used as a gatekeeping strategy to test the primary family of endpoints, the Bonferroni
1394   method and some other single-step methods (such as the Dunnett's test, which is not covered in

1395　this document) have an important property of preserving some alpha for testing the secondary
1396　endpoint family when at least one of the endpoints in the primary family is statistically
1397　significant. In the Bonferroni method, the endpoint-specific alpha from each test that
1398　successfully rejected that null hypothesis is summed and becomes the alpha available to the
1399　secondary endpoint family. For example, in the equally weighted Bonferroni method, when
1400　there are two endpoints in the primary family, the unused alpha available for tests of hypotheses
1401　in the secondary family can be 0.05, 0.025, or 0, depending, respectively, on whether both, one,
1402　or none of the primary endpoint tests rejected their respective null hypotheses.
1403
1404　The conventional Holm and Hochberg methods, however (see sections IV.C.2 and IV.C.3), do
1405　not have this property. These methods pass alpha from the primary family to the secondary
1406　family only when all of the null hypotheses in the primary family are rejected. These two
1407　methods give better power on recycling all alpha within the family and releasing it only when all
1408　hypotheses in that family are rejected. Inappropriately proceeding as if there is some preserved
1409　alpha when a study fails to reject one or more of the primary hypotheses will result in an inflated
1410　overall Type I error rate.
1411
1412　There are, however, procedures called the truncated Holm and the truncated Hochberg that can
1413　be used when there is a desire to have the power advantage of the conventional Holm or
1414　Hochberg procedures but also to have some alpha available for testing the secondary endpoint
1415　family if at least one of the primary endpoints is successful. In a truncated Holm or Hochberg
1416　procedure, some portion of the unused alpha from each step is reserved for passing to the
1417　secondary endpoint family. The truncated Holm procedure and the truncated Hochberg
1418　procedures are hybrids of their conventional forms and the Bonferroni method. As a
1419　consequence, the endpoint-specific alpha for each successive test of hypothesis of the primary
1420　endpoints after the first is not as large as in the conventional Holm or the conventional Hochberg
1421　procedure. In either of these approaches, of course, if all of the individual endpoint tests of
1422　hypotheses in the primary endpoint family successfully reject the null hypothesis, the full alpha
1423　of 0.05 is available for the secondary endpoint family. The amount of reserved alpha from the
1424　successive tests should be chosen carefully, as the choice creates a balance between decreasing
1425　study power for the endpoints in the primary family and the guarantee (if at least the first test
1426　rejects the null hypothesis) of some power to test the secondary endpoint family. The following
1427　example illustrates these two procedures for a primary family with three endpoints.
1428
1429　Consider treatment versus control comparisons for three endpoints in the primary family with the
1430　control of alpha at the 0.05 level. The endpoint-specific alpha levels for the conventional Holm
1431　for this case are 0.05/3, 0.05/2, and 0.05 (see section IV.C.2), and those by the equally weighted
1432　Bonferroni method are 0.05/3, the same for each comparison (see section IV.C.1). The endpoint-
1433　specific alpha levels for the truncated Holm are then constructed by combining the endpoint-
1434　specific alpha levels of the two methods with a "truncation fraction" of f, whose value between
1435　zero and one is selected in advance. The following calculations illustrate this combination using
1436　$f = 1/2$; the multipliers with f are the endpoint-specific alpha levels for the conventional Holm and
1437　those with (1-f) are by the equally weighted Bonferroni method.
1438

1439
$$\alpha_1 = \frac{0.05}{3}f + \frac{0.05}{3}(1-f) = \frac{0.05}{3}\cdot\frac{1}{2} + \frac{0.05}{3}(1-\frac{1}{2}) = 0.0167$$

1440
$$\alpha_2 = \dfrac{\dfrac{0.05}{2}f + \dfrac{0.05}{3}(1-f)}{} = \dfrac{\dfrac{0.05}{2}\cdot\dfrac{1}{2} + \dfrac{0.05}{3}(1-\dfrac{1}{2})}{} = 0.0208$$

1441
$$\alpha_3 = \dfrac{\dfrac{0.05}{1}f + \dfrac{0.05}{3}(1-f)}{} = \dfrac{\dfrac{0.05}{1}\cdot\dfrac{1}{2} + \dfrac{0.05}{3}(1-\dfrac{1}{2})}{} = 0.0333$$

1442
1443 Thus, for this particular case, when the value of $f = 1/2$, the first test for the truncated Holm test
1444 is performed at $\alpha_1 = 0.0167$, which is the same for the conventional Holm test. However, the
1445 second test, after the first test is successful, is performed at level $\alpha_2 = 0.0208$, and the third test,
1446 after the first two tests are successful, is at level $\alpha_3 = 0.0333$. The unused alpha levels for passing
1447 to the secondary family are calculated as:
1448
1449   i.   Unused alpha = 0.05, if all three tests are successful;
1450   ii.   Unused alpha = $(0.05 - \alpha_3) = 0.05 - 0.0333 = 0.0167$, if the first two tests are successful,
1451   but the last one is not;
1452   iii.   Unused alpha = $(0.05 - 2\alpha_2) = 0.05 - 2(0.0208) = 0.0084$, if the first test is successful,
1453   but the other two tests are not.
1454
1455 For the truncated Hochberg, alpha levels α_1, α_2, and α_3 are the same as those for the truncated
1456 Holm, except that for the truncated Hochberg, the first test starts with the largest p-value (i.e.,
1457 largest of the three endpoint treatment-to-control comparison p-values) at level $\alpha_3 = 0.0333$. If
1458 this first test is successful, then the other two tests are also considered successful, and alpha of
1459 0.05 passes to the secondary family. However, if the first test is not successful, then the second
1460 test with second-largest p-value is at level $\alpha_2 = 0.0208$. If this second test is successful, then the
1461 remaining last test is also considered successful, and alpha of 0.0167 passes to the secondary
1462 family. However, if this second test is not successful, then the last test with the smallest p-value
1463 is at level $\alpha_1 = 0.0167$, and if that test is successful, then alpha of 0.0084 passes to the secondary
1464 family. This illustration is with $f = 1/2$. Similar calculations would follow for different values of
1465 f.
1466
1467   9.   *Multi-Branched Gatekeeping Procedures*
1468
1469 Some multiplicity problems are multidimensional. One dimension may correspond to multiple
1470 endpoints, a second to multiple-dose groups (that have each of those endpoints tested), and yet
1471 another dimension to multiple hypotheses regarding an endpoint, such as non-inferiority and
1472 superiority tests (for each dose and each endpoint). Each individual hypothesis to test pertains to
1473 one particular endpoint, dose, and analysis objective. The total number of hypotheses is the
1474 product of the number of options within each dimension and can become large, even when there
1475 are only two or three options for each dimension.
1476
1477 The multiple sources of multiplicity create the potential for multiple pathways of testing the
1478 hypotheses. For example, if the goal of a study is to demonstrate non-inferiority as well as
1479 superiority, a single path of sequential tests is preferred. After demonstrating non-inferiority on
1480 the endpoint, it is possible to then test for superiority at an unadjusted alpha. In a fixed-sequence
1481 (unbranched) approach, it would also be appropriate to analyze a second endpoint for non-
1482 inferiority at the same alpha after the first endpoint is successfully shown to be non-inferior.

1483   Suppose, however, that one wants to carry out both of these analyses after showing non-
1484   inferiority for the first endpoint. The testing path now branches into two paths from this initial
1485   test, i.e., testing superiority for the first endpoint and non-inferiority for the second endpoint.
1486   There is a choice of statistical adjustments to apply in this setting.
1487
1488   Treating the hypotheses as independent and applying a simple method such as Bonferroni leads
1489   to testing these hypotheses at small alpha levels, and consequently a very large study may be
1490   necessary to ensure good study power. Alternatively, applying a fixed-sequence method may
1491   lead to many endpoint tests being disallowed because the optimal sequence for testing is usually
1492   not prospectively determinable. The multi-branched gatekeeping procedure can address
1493   multiplicity problems of this multi-dimensional type. The multi-branched gatekeeping procedure
1494   allows for ordering the sequence of testing with the option of testing of more than one endpoint
1495   if a preceding test is successful. When there are multiple levels of this sequential hierarchy, and
1496   branching is applied at several of the steps, the possible paths of endpoint testing become a
1497   complex, multi-branched structure.
1498
1499   As a simple illustration (Figure 1), consider a clinical trial that compares a treatment to control
1500   on two primary endpoints (endpoint one and endpoint two) to determine first whether the
1501   treatment is non-inferior to the control for at least one endpoint. If, for either of the two
1502   endpoints, the treatment is found non-inferior to the control, there is also a desire to test whether
1503   it is superior to control for that endpoint. The analytic plan for the trial thus sets the following
1504   logical restrictions:
1505
1506   i.   Test endpoint two only after non-inferiority for endpoint one is first established.
1507   ii.   Test for superiority on an endpoint only after non-inferiority for that endpoint is first
1508   concluded.
1509
1510   The following diagram shows the decision structure of the test strategy. In this diagram, each
1511   block (or node) states the null hypothesis that it tests.
1512
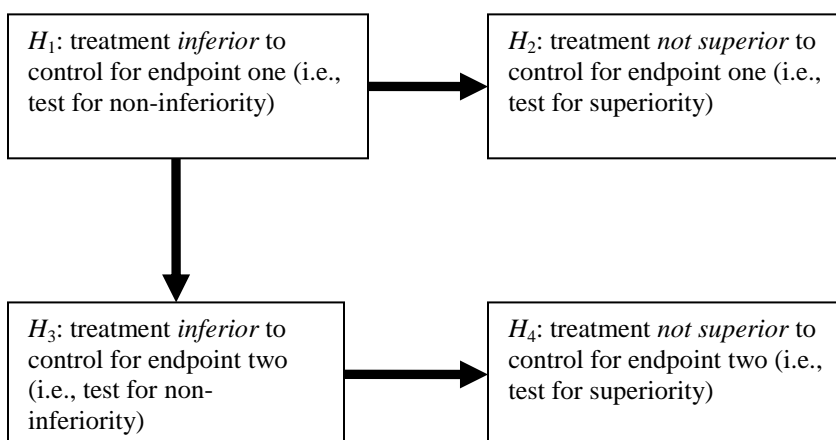
1513
1514
1515   **Figure 1:** Example of a flow diagram for non-inferiority and superiority tests for endpoints one and two of a trial
1516   with logical restrictions: in order to test for superiority for endpoint one and/or two, one must first establish non-
1517   inferiority for that endpoint.
1518

1519 Thus, the above test strategy has a two-dimensional hierarchical structure, one dimension for the
1520 two different endpoints and the other for the non-inferiority and superiority tests, with the logical
1521 restrictions as stated above. A different study might have three dimensions, two endpoints to be
1522 tested at two dose levels (along with a control group) with non-inferiority and superiority tests on
1523 each endpoint, having restrictions, e.g., that the lower dose can be tested after a success on the
1524 higher dose, and superiority on an endpoint can be tested after non-inferiority has been shown.
1525
1526 For the test strategy in Figure 1, one may, inappropriately, test each hypothesis at the same
1527 significance level (e.g., $\alpha = 0.05$), reasoning that the tests for non-inferiority for the two
1528 endpoints follow a sequential order, allowing passing along the full alpha; and that the test for
1529 superiority for each endpoint follows naturally after non-inferiority for it is first demonstrated.
1530 This approach, however, is likely to inflate the overall Type I error rate, because in Figure 1, the
1531 testing path (sequence) after the node at H_1 splits into two branches; one goes on to test for H_2
1532 and the other to test for H_3. Consequently, once the trial concludes non-inferiority of the
1533 treatment to control for endpoint one, erroneous conclusions for tests of H_2 and H_3 can occur in
1534 multiple ways; that is, either H_2 is erroneously rejected, or H_3 is erroneously rejected, or both H_2
1535 and H_3 are erroneously rejected. If each of these separate hypotheses were to be tested at the
1536 0.05 level, this would obviously lead to Type I error rate inflation. As another illustration of
1537 Type I error rate inflation, suppose that in reality the treatment is non-inferior to control for both
1538 endpoints but is not superior to control for either endpoint. In this scenario, the testing scheme
1539 (without alpha adjustments) can conclude superiority of the treatment to control in multiple
1540 ways, i.e., the treatment is superior to control for either endpoint one or endpoint two, or for both
1541 endpoints.
1542
1543 It is possible to deal with this problem using the Bonferroni-based gatekeeping method by
1544 grouping the hypotheses as follows:
1545
1546 • Group one includes only H_1 (the test of non-inferiority for endpoint one)
1547 • Group two includes H_2 (the test of superiority for endpoint one) and H_3 (the test of non-
1548   inferiority for endpoint two)
1549 • Group three includes only H_4 (the test of superiority for endpoint two).
1550
1551 The procedure would begin with the test of the single hypothesis H_1 in group one at the level
1552 intended for the study-wise overall Type I error rate (e.g., $\alpha = 0.05$). Group one serves as a
1553 gatekeeper for group two. Therefore, once the result for H_1 is significant at level α (i.e., the
1554 treatment is non-inferior to control for endpoint one at level α), testing proceeds to the
1555 hypotheses H_2 and H_3 in group two with the alpha that was not used within family one, which in
1556 this case would be the overall study alpha.
1557
1558 The test of H_2 and H_3 in family two can use the Bonferroni method at the endpoint-specific alpha
1559 of 0.025 for each test according to the Bonferroni-based gatekeeping method. The standard
1560 Holm procedure is not considered here for the reason discussed in sections IV.C.2 and IV.C.8.
1561 Dividing the available alpha between the two endpoints will reduce study power for these
1562 endpoints (or necessitate an increased sample size to maintain study power), making it more
1563 difficult for the study to succeed on these endpoints; but it is necessary to maintain control of
1564 Type I error rate.

1565
1566   Therefore, if both H_2 and H_3 are rejected, H_4 is tested at $\alpha = 0.05$. However, if only H_3 is
1567   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
1568   at $\alpha = 0.025$ in accord with the plan, but this would be illogical because if endpoint two failed to
1569   show non-inferiority (H_3), superiority could not have occurred.
1570
1571   When there are three or more dimensions and multiple branch points, planning the sequence of
1572   testing becomes complex and difficult to describe in the manner illustrated here. In these
1573   situations, the graphical approach to displaying and evaluating analysis paths (Appendix A) can
1574   be valuable.
1575
1576   *10.   Resampling-Based, Multiple-Testing Procedures*
1577
1578   When there is correlation among multiple endpoints, resampling is one general statistical
1579   approach that can provide more power than the methods described above to detect a true
1580   treatment effect while maintaining control of the overall Type I error rate, and the power
1581   increases as the correlation increases. With these methods, a distribution of the possible test-
1582   statistic values under the null hypothesis is generated based upon the observed data of the trial.
1583   This data-based distribution is then used to find the p-value of the observed study result instead
1584   of using a theoretical distribution of the test statistics (e.g., a normal distribution of Z-scores, or a
1585   t-distribution for t-scores) as with most other methods.
1586
1587   Resampling methods include the bootstrap and permutation approaches for multiple endpoints
1588   and require few, albeit important, assumptions about the true distribution of the endpoints. There
1589   are, however, some drawbacks to these methods. The important assumptions are generally
1590   difficult to verify, particularly for small study sample sizes. These methods, consequently,
1591   usually require large study sample sizes (particularly bootstrap methods) and often require
1592   simulations to ensure the data-based distribution of the test statistics from the limited trial data is
1593   applicable and to ensure adequate Type I error rate control. Inflation of the Type I error rate may
1594   occur, for example, if the shape of the data distribution is different between the treatment groups
1595   being compared.
1596
1597   There is at present little experience with these methods in drug development clinical trials.
1598   Because of this, resampling methods are not recommended as primary analysis methods for
1599   adequate and well-controlled trials in drug development. It may, however, be useful and
1600   instructive to compare the results of resampling methods with those obtained using conventional
1601   methods in order to gain experience with and understanding of resampling methods' properties,
1602   advantages, and limitations.
1603
1604
1605   **V.   CONCLUSION**
1606
1607   The chance of making a false positive conclusion, concluding that a drug has a beneficial effect
1608   when it does not, is of primary concern to FDA. The widely accepted standard is to control the
1609   chance of coming to a false positive conclusion (Type I error probability) about a drug's effects
1610   to less than 2.5 percent (1 in 40 chance). As the number of endpoints or analyses increases, the

1611   probability of making a false positive conclusion can increase well beyond the 2.5 percent
1612   standard. Multiplicity adjustments, as described in this guidance, provide a means for
1613   controlling Type I error when there are multiple analyses of the drug's effects. There are many
1614   strategies and/or choices of methods that may be used, as appropriate, as described in this
1615   guidance. Each of these methods has advantages and disadvantages and the selection of suitable
1616   strategies and methods is a challenge to be addressed at the study-planning stage. Statistical
1617   expertise should be enlisted to help choose the most appropriate approach. Failure to
1618   appropriately control the Type I error rate can lead to false positive conclusions; this guidance is
1619   intended to clarify when and how multiplicity due to multiple endpoints should be managed to
1620   avoid reaching such false conclusions.
1621

1622 **GENERAL REFERENCES**
1623

1624 Alosh M, Bretz F, Huque MF. Advanced multiplicity adjustment methods in clinical trials.
1625 *Statistics in Medicine* 2014; **33**(4): 693-713.

1626 Bauer P. Multiple testing in clinical trials. *Statistics in Medicine* 1991; **10**: 871-890.

1627 Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*, CRC Press (Taylor & Francis
1628 Group), Chapman and Hall, 2010.

1629 Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective
1630 multiple test procedures. *Statistics in Medicine* 2009; **28**: 586-604.

1631 Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for
1632 multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests.
1633 *Biometrical Journal* 2011; **53**(6): 894-913.

1634 Chi GYH. Some issues with composite endpoints in clinical trials. *Fundamental & Clinical
1635 Pharmacology* 2005; **19**: 609-619.

1636 CPMP/EWP/908/99. Points to consider on multiplicity issues in clinical trials. September 2002;
1637 http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500
1638 003640.pdf.
1639
1640 Dmitrienko A, Tamhane AC, Bretz F. *Multiple testing problems in pharmaceutical statistics*,
1641 CRC Press (Taylor & Francis Group), Chapman & Hall/CRC Biostatistics Series, 2010.

1642 Dmitrienko A, D'Agostino RB, Huque MF. Key multiplicity issues in clinical drug
1643 development. *Statistics in Medicine* 2013; **32**: 1079–1111.

1644 Dmitrienko A, D'Agostino RB. Tutorial in Biostatistics: Traditional multiplicity adjustment
1645 methods in clinical trials. *Statistics in Medicine* 2013; **32**(29): 5172-5218.

1646 Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*
1647 1988; **75**: 800-802.

1648 Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. John Wiley & Sons, New York,
1649 1987.

1650 Holm SA. A simple sequentially rejective multiple test procedure. *Scandanavian Journal of
1651 Statistics* 1979; **6:** 65-70.

1652 Hommel G, Bretz F, Maurer W. Multiple hypotheses testing based on ordered p values — a
1653 historical survey with applications to medical research. *Journal of Biopharmaceutical Statistics*
1654 2011; **21**(4): 595-609.

1655 Hung HMJ, Wang SJ. Challenges to multiple testing in clinical trials. *Biometrical Journal* 2010;
1656 **52**(6): 747-756.

1657 Huque MF. Validity of the Hochberg procedure revisited for clinical trial applications. *Statistics*
1658 *in Medicine* 2015, (wileyonlinelibrary.com) DOI: 10.1002/sim.6617.

1659 Huque MF, Alosh M, Bhore R. Addressing multiplicity issues of a composite endpoint and its
1660 components in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **21**: 610-634.

1661 Huque MF, Dmitrienko A, D'Agostino RB. Multiplicity issues in clinical trials with multiple
1662 objectives. *Statistics in Biopharmaceutical Research* 2013; 5(4): 321-337.

1663 Lubsen J, Kirwan BA. Combined endpoints: can we use them? *Statistics in Medicine* 2002; **21**:
1664 2959–2970.

1665 Moye LA. *Multiple Analyses in Clinical Trials*. Springer-Verlag, New York, 2003.

1666 O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not
1667 demonstrate clear statistical significance. *Controlled Clinicial Trials* 1997; **18**: 550-556.

1668 Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of
1669 composite endpoints in clinical trials based on clinical priorities. *European Heart Journal* 2012;
1670 **33**: 176–182.

1671 Sarkar S, Chang CK. Simes' method for multiple hypotheses testing with positively dependent
1672 test statistics. *Journal of the American Statistical Association* 1997; **92**: 1601-1608.

1673 Westfall PH, Tobias RD, Rom D, Wolfinger RD, HochbergY. *Multiple Comparisons and*
1674 *Multiple Tests Using the SAS® System*, SAS Institute Inc.: Cary, NC, USA, 1999.

1675 Westfall PH, Young SS. *Resampling Based Multiple Testing: Examples and Methods for P-*
1676 *value Adjustment*. John Wiley & Sons, Inc. New York, 1993.

1677 Wiens BL. A fixed sequence Bonferroni procedure for testing multiple endpoints.
1678 *Pharmaceutical Statistics* 2003; **2**: 211-215.

1679
1680 **REFERENCES TO EXAMPLES**

1681 Brenner BM, Cooper ME, de Zeeuw D, Keane WF, Mitch WE, Parving H-H, Remuzzi G,
1682 Snapinn SM, Zhang Z, and Shahinfar S, for the RENAAL Study Investigators. Effects of
1683 Losartan on Renal and Cardiovascular Outcomes in Patients with Type 2 Diabetes and
1684 Nephropathy. *New England Journal of Medicine* 2001; 345:861-869.
1685
1686 Dahlöf G, Devereux RB, Kjeldsen SE, Julius S, Beevers G, de Faire U, Fyhrquist F, Ibsen H,
1687 Kristiansson K, Lederballe-Pedersen O, Lindholm LH, Nieminen MS, Omvik P, Oparil S, Wedel

1688   H: LIFE Study Group. Cardiovascular morbidity and mortality in the Losartan Intervention For
1689   Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. *Lancet*
1690   2002; 359(9311): 995-1003.
1691

APPENDIX: THE GRAPHICAL APPROACH

1692
1693
1694   A graphical approach is available for developing and evaluating hierarchal multiple analysis
1695   strategies. This approach provides a means for specifying, communicating, and assessing
1696   different hypothesis testing strategies, but is not by itself an additional method for addressing
1697   multiplicity (such as those described in section IV). Instead, the graphical approach is a means
1698   of depicting a strategy consisting of the previously described Bonferroni-based sequential
1699   methods, such as fixed-sequence, fallback type, and gatekeeping procedures. This approach
1700   illustrates differences in endpoint importance as well as the relationships among the endpoints by
1701   mapping onto a test strategy that ensures control of the Type I error rate and aids in creating and
1702   evaluating alternative test strategies. This technique will be most helpful when the analysis plan
1703   is complex due to splitting of the overall alpha among several endpoints (either initially or after a
1704   particular endpoint has been successful), particularly if there is a desire to have a second chance
1705   for an endpoint that was not statistically significant at the initially assigned endpoint-specific
1706   alpha, but can receive pass-along alpha from a different endpoint that was successful (the loop-
1707   back feature described below). This situation may occur when complex testing strategies are
1708   being considered because of intricate endpoint relationships and differing endpoint importance.
1709
1710   Graphical displays of complex analysis strategies can aid in clearly describing and assessing the
1711   proposed plan by displaying all the logical relationships among endpoint tests of hypotheses. In
1712   addition, simple modifications of the initial graph can easily create different variations of a test
1713   strategy, aiding comparison among the variations. The graphical approach can be useful in trial
1714   design to identify a test scheme that is suitably tailored to the objectives of the trial.
1715
1716   *Basics of the Graphical Approach: Use of vertex (node) and path (order or direction)*
1717
1718   In the graphical approach, the testing strategy is defined by a figure that shows each of the
1719   hypotheses (H_1, H_2,..., H_m) located at a vertex (or node, a junction of testing order paths), and
1720   depicts the test order paths by lines (with the direction of the path indicated by an arrowhead)
1721   connecting the hypotheses. Each vertex (hypothesis) is allocated an initial amount of alpha,
1722   which we call here the "endpoint-specific alpha" (with the understanding that a test of an
1723   endpoint is associated with a test of a hypothesis, and vice versa). A key requirement is that the
1724   sum of all of the endpoint-specific alpha levels is equal to the total alpha level available for the
1725   study (the study-wise Type I error rate). An exception can occur if one designates two or more
1726   hypotheses as a co-endpoint group, so that the same endpoint-specific alpha is applied to all tests
1727   in that group.
1728
1729   Each test order path is also assigned a value between 0 and 1, called a weight for that path and
1730   shown above the arrow, which indicates the fraction of the preserved alpha to be shifted along
1731   that path to the receiving hypothesis, when the hypothesis at the tail end of the path is successful
1732   (i.e., is rejected). The sum of the weights across all the paths leaving a vertex must be 1.0, so
1733   that the entire preserved alpha is used in testing subsequent hypotheses.
1734
1735   All study hypotheses that are intended to potentially provide firm conclusions of efficacy are
1736   shown in the graph. With this technique there is no need to explicitly designate hypotheses as
1737   part of the primary or secondary endpoint families; more nuanced hierarchies are able to be

1738   achieved based on the initial allocation of the endpoint-specific alpha and the division of passed-
1739   forward alpha among the test paths leaving each vertex. Clearly, the hypotheses that receive an
1740   initial endpoint-specific alpha allocation of 0.0 will often be those regarded as of lesser
1741   importance, which is implicitly similar to designating the associated endpoint as a secondary
1742   endpoint.
1743
1744   Adhering to the principles outlined in prior sections of this guidance, when an endpoint test is
1745   successful in rejecting the corresponding null hypothesis, that endpoint-specific alpha can be
1746   passed on to the next test indicated by the arrow, and will be divided among several subsequent
1747   hypotheses when there are several paths leaving that vertex. This shift of alpha occurs only
1748   when the test result for the hypothesis associated with a vertex at the arrow's tail is significant.
1749   Thus, as with the simple fallback method, the actual endpoint-specific alpha used in an endpoint
1750   test cannot be determined until the study results are complete and hypothesis testing begins; the
1751   sequential test determines which vertices are associated with alpha levels that can be passed
1752   along for accumulation in the subsequent test and which are not.
1753
1754   Several examples of the graphical method follow to help illustrate the concept, construction,
1755   interpretation, and application of these diagrams. The first several of these examples are simple
1756   cases where the graphical approach is no more useful than a nondiagramatic (written text)
1757   description, but where the principles of the approach can be more clearly illustrated.
1758
1759   ***Fixed-Sequence Method***
1760
1761   The fixed-sequence testing strategy (section IV.C.5), shown in Figure A1, illustrates a simple
1762   case of the graphical method with three hypotheses. In this scheme, the endpoints (hypotheses)
1763   are ordered. Testing begins with the first endpoint at the full alpha level, and continues through
1764   the sequence only until an endpoint is not statistically significant. This diagram shows that the
1765   endpoint-specific alpha levels associated with hypotheses H_1, H_2, and H_3 are set in the beginning
1766   as α, 0, and 0. Arrows indicate the sequence of testing, and if the test is successful, the full alpha
1767   is shifted along to the next test. Consequently, if null hypothesis H_1 is successfully rejected, the
1768   endpoint-specific alpha level for H_2 becomes $0 + 1 \times \alpha = \alpha$, which allows testing of H_2 at level α.
1769   However, if the test of H_1 is unsuccessful, there is no pre-assigned non-zero alpha for H_2 to allow
1770   testing of H_2, so the testing stops.
1771
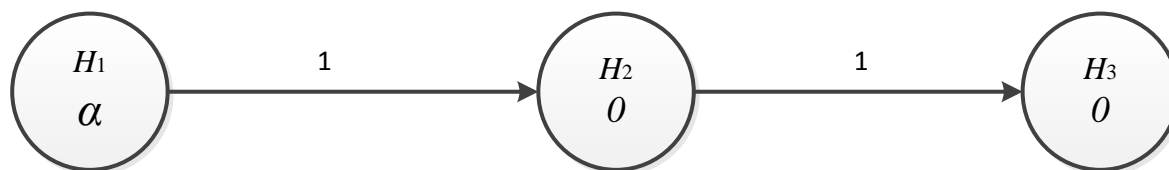
1772
1773   **Figure A1**: Graphical illustration of the fixed-sequence testing with three hypotheses.
1774
1775

1776 ***Loop-Back Feature to Indicate Two-Way Potential for Alpha Passing***
1777
1778 Another valuable feature of the graphical method occurs when the available alpha level is split
1779 between two or more endpoints into endpoint-specific alphas levels; these diagrams can easily
1780 illustrate the potential for loop-back passing of endpoint-specific alpha. If a hypothesis is not
1781 rejected at its endpoint-specific alpha level, but a different hypothesis is, then the unused
1782 endpoint-specific alpha from the rejected second hypothesis can be directed to loop back to the
1783 first hypothesis, which is then re-tested at the higher alpha level. Thus, in Figure A2, if assigned
1784 endpoint-specific alpha levels for testing H_1 and H_2 are $\alpha_1 = 0.04$ and $\alpha_2 = 0.01$, respectively, and
1785 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
1786 re-testing at the higher level of $0.04 + 0.01 = 0.05$. Without the loop-back from H_2 to H_1, this
1787 would simply be the fallback method (described in section IV.C.6).
1788
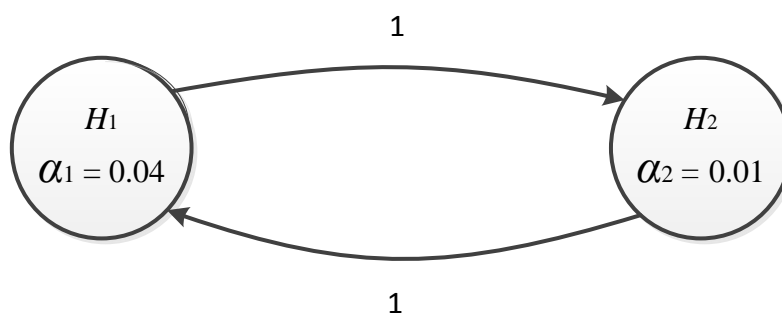
1789
1790
1791 **Figure A2**: Graphical illustration of the loop back passing of endpoint-specific alpha from H_2 to
1792 H_1.
1793
1794 The Holm procedure (section IV.C.2) is a specific case of tests for two hypotheses with a loop-
1795 back feature where the graphical method enables a simple depiction of the procedure and its
1796 rationale. The Holm procedure directs that the first step is to test the smaller p-value at endpoint-
1797 specific alpha $= \alpha/2$ and, only if successful, proceed to test the larger p-value at the level α (e.g.,
1798 0.05). Because the Holm procedure splits alpha evenly in half, if the test of hypothesis with the
1799 smaller p-value was not significant, it is clear that the test with the larger p-value will also fail to
1800 be significant; performing that comparison is unnecessary. The diagram for the Holm procedure
1801 (Figure A3), shows two vertices and associated endpoint-specific alpha levels of $\alpha_1 = 0.025$ and
1802 $\alpha_2 = 0.025$, respectively, satisfying the requirement for total alpha $= 0.05$. The two arrows show
1803 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
1804 endpoint-specific alpha of 0.025 is shifted entirely to the other hypothesis, and added to the
1805 endpoint-specific alpha already allocated for that hypothesis to provide a net alpha of 0.05.
1806 Because either hypothesis might be tested first, the diagram shows a loop-back configuration.
1807
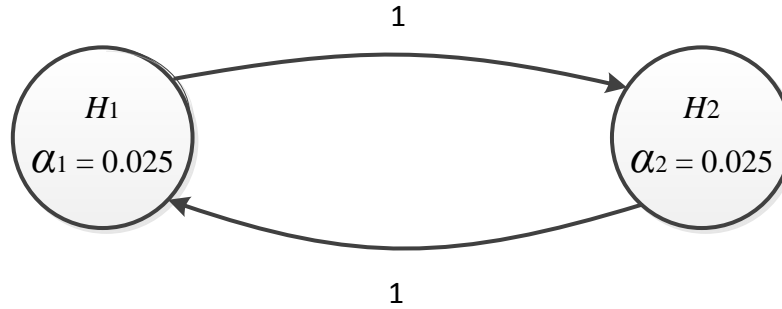
Figure A3: Graphical illustration of the Holm procedure with two hypotheses.

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
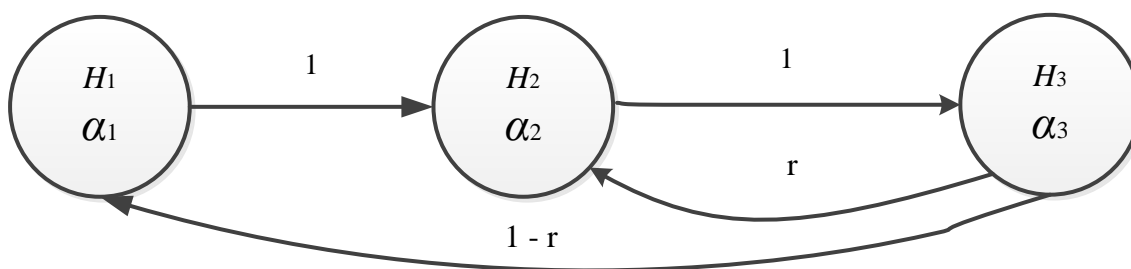
1847 earlier endpoints might most benefit from this passing-back of alpha). Figure A4(b) shows this
1848 procedure with the fraction r of this un-used alpha shifted to H_2 and the remaining fraction 1- r of
1849 this alpha shifted to H_1. The value of r should be prospectively specified in the study analysis
1850 plan.
1851
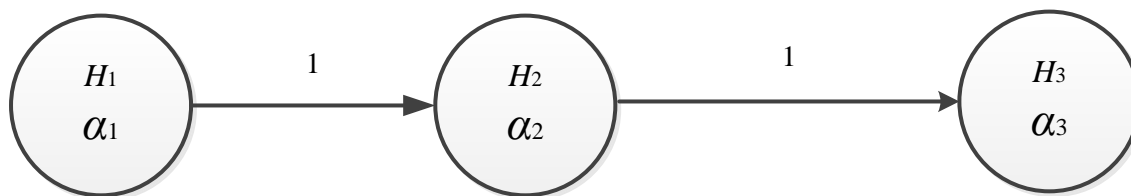
(a)

1852
1853

(b)

1854
1855
1856 **Figure A4:** Fallback (a) and improved fallback (b) procedures.
1857

1858
1859 ***Progressive Updating of the Diagram When Hypotheses Are Successfully Rejected***
1860
1861 The graphical approach guides the hierarchical testing of multiple hypotheses through continual
1862 updating of the initial graph whenever a hypothesis is successfully rejected. The initial graph
1863 represents the full testing strategy (with all hypotheses). Each new graph shows the progression
1864 of the testing strategy by eliminating hypotheses that have been rejected and retaining those yet
1865 to be tested or re-tested.
1866
1867 When there is a desire to consider analysis strategies with complex division of alpha, the
1868 graphical method and progressive updating of the diagram can aid in understanding the
1869 implication of the different strategies for a variety of different hypothetical scenarios. This
1870 progressive updating can aid in selecting which specific strategy to select for the final study
1871 statistical analysis plan.
1872
1873 Figure A5 is an example of how the graphical method aids in formulating the testing of three
1874 hypotheses H_1, H_2, and H_3 and illustrates the updating of the diagram when a test of hypothesis is

1875 successful. For this example, the analysis plan designated two hypotheses, H_1 and H_2, to be of
1876 prime importance (i.e., primary endpoints), and H_3 (the secondary endpoint) is tested only if the
1877 test results for H_1 and H_2 are both significant. Assume that it is desired to always be able to test
1878 both H_1 and H_2 (i.e., a willingness to split the available alpha between them), but that if either H_1
1879 or H_2 is successfully rejected, the alpha level of that test would be passed to the other hypothesis
1880 if needed, so that it can be tested at the maximal possible alpha level (i.e., the fallback method is
1881 specified for the two important endpoints, with α_1 assigned to begin testing on H_1, and α_2
1882 reserved as the minimum that will be available for testing H_2). Thus, as shown in Figure A5 (a),
1883 if the test result for H_1 is significant, then its endpoint-specific alpha level α_1 is passed to H_2, so
1884 that H_2 is tested at an endpoint-specific alpha level of $\alpha_1 + \alpha_2 = \alpha$. On the other hand, if the test
1885 result for H_1 is not significant, H_2 is still tested with the reserved α_2. In this case, however, if the
1886 test result for H_2 is significant, the alpha level of α_2 is recycled back to re-testing of H_1 at level α_1
1887 $+ \alpha_2 = \alpha$. Note that the graphical method aids in communicating that the re-testing of H_1 at an
1888 increased endpoint-specific alpha is part of the prospective analytic plan.
1889
1890 The intended analysis, however, is that if, and only if, these tests of hypotheses (including
1891 potential re-test with passed alpha) have successfully rejected H_1 and H_2, then the full available
1892 alpha would be passed to H_3. This conditional passing of alpha is depicted by a path from H_2 to
1893 H_3 with weight ε. At the start (before any testing of any hypothesis) ε is set to a negligible
1894 amount. Because of this, even though there is a path from H_2 to H_3, when H_1 has not yet been
1895 successfully rejected, essentially all of α_2 will be passed back to H_1 as the priority over H_3. This
1896 scheme will eventually allow for meaningful testing of H_3 if appropriate according to the
1897 sequentially updated diagrams.
1898
1899 Figure A5 (b) shows the updated graph when the result for H_1 in Figure A5 (a) is significant at
1900 level α_1 and prior to testing H_2 at the now accumulated endpoint-specific alpha of $\alpha_1 + \alpha_2$ (which
1901 would be equal to the total alpha for the study in this case). Note that the weight on the path
1902 from H_2 to H_3 is now set to 1. This occurs because diagram updating is done when a test of
1903 hypothesis is significant. The process of diagram updating first passes along the retained alpha
1904 from the successful hypothesis (vertex) according to the weights on the arrows leaving that
1905 vertex. That vertex is then eliminated from the diagram and a new diagram is constructed by
1906 connecting all the incoming paths (arrows) to all outgoing paths (tails) of the now deleted vertex,
1907 and adjusting the pathway weights. The new weights on the new paths are determined based on
1908 the relative weights of each previous part of the new path. The essential principle of
1909 readjustment of the pathway weights is that the sum of the weights on the outgoing paths from
1910 each vertex must be 1.0. This rule causes the weight on the path from H_2 to H_3 to become 1
1911 (from the prior negligible fraction ε) because it is the only remaining path leaving H_2. In some
1912 strategies, a newly created connection path arising from elimination of a successful vertex will
1913 duplicate a preexisting direct connection between two vertices; in this case the weights of the
1914 duplicate paths are combined and drawn as a single path.
1915
1916 Continuing with the example depicted in Figure A5, if H_1 is not initially significant and H_2 is
1917 significant at level α_2, Figure A5 (c) shows the updated diagram prior to re-testing H_1 at the now
1918 accumulated endpoint-specific alpha. The vertex for H_2 was eliminated from the updated
1919 diagram, and the direct path from H_1 to H_3 is displayed. Both Figures A5 (b) and A5 (c) indicate

1920   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
1921   significant, but that no alpha is passed to H_3 unless both H_1 and H_2 were significant.

(a)

(b)

(c)
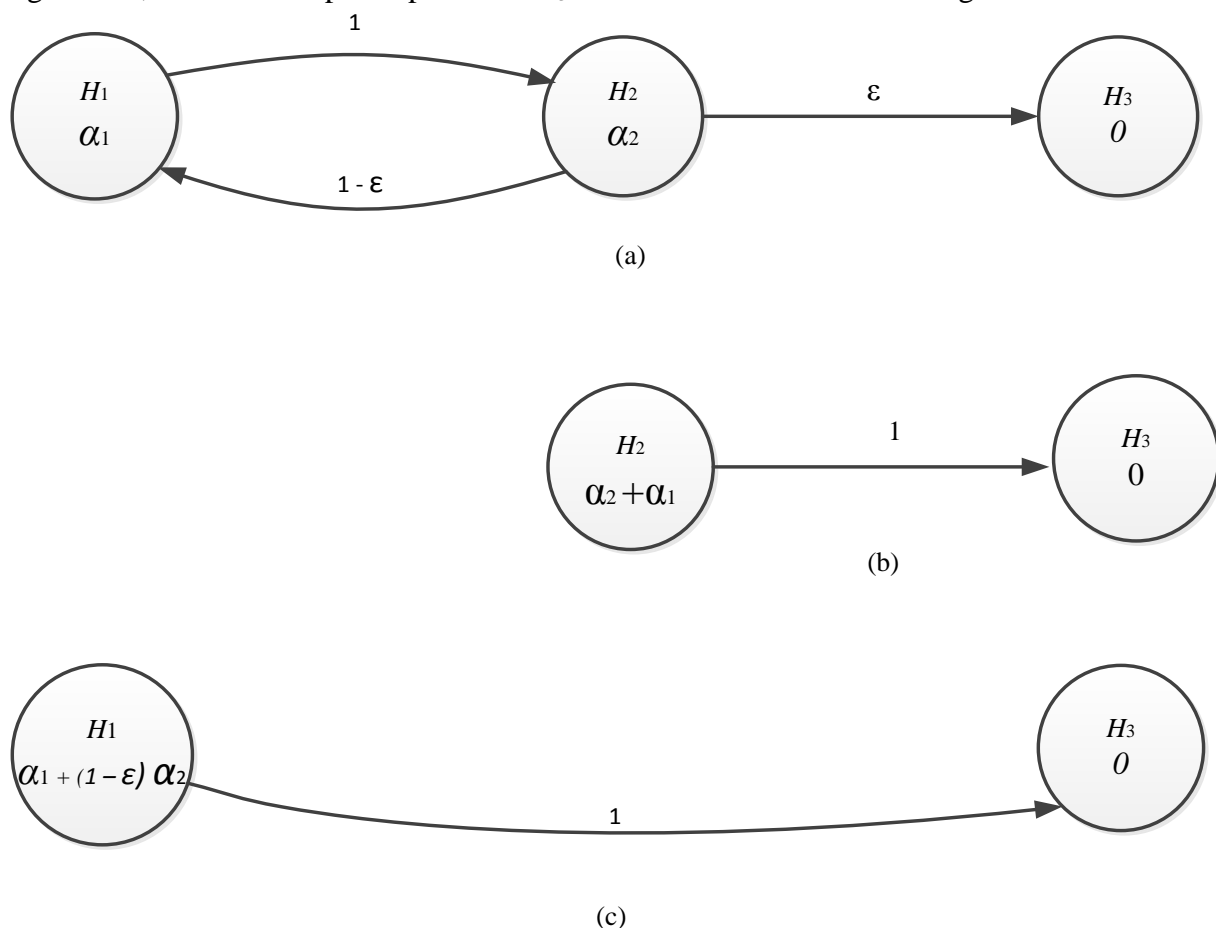
1922
1923   **Figure A5**: Graphical illustration of the fallback procedure applied to three hypotheses when the
1924   first two hypotheses are most important and the third hypothesis is tested only when both of the
1925   first two hypotheses are significant.
1926   (a) The initial diagram shows all hypotheses and paths. The notation ε indicates a positive
1927   number close to zero. This convention indicates the potential to pass alpha to H_3, but only
1928   if it is not necessary to pass alpha from H_2 to H_1 (see text for explanation).
1929   (b) The updated diagram shows the case where only H_1 was tested and shown to be
1930   statistically significant.
1931   (c) The updated diagram shows the case where H_2 was the first hypothesis to be statistically
1932   significant at the initially allocated endpoint-specific alpha.
1933

1934   A detailed algorithm for iteratively updating the graph when a test is found significant is
1935   illustrated with the final example. Updating of a graph involves determining new endpoint-
1936   specific alpha levels and path weights based on satisfying the conditions that (1) the sum of all
1937   endpoint-specific alpha levels equals α and (2) the sum of all weights on outgoing arrows from a
1938   vertex to others equals 1.0.
1939

1940 The case of three hypotheses with fixed weights on the paths between the hypotheses will be
1941 used to illustrate the algorithm (Figure A6 (a)). Suppose that hypothesis H_3 is rejected. The
1942 graph needs to be updated to remove this hypothesis and retain hypotheses H_1 and H_2.
1943 Calculations for this are as follows:
1944
1945   1. New alpha level at H_1 = old alpha level at H_1 + w_{31} × (the alpha level at H_3) = $\alpha/3$ + (1) ×
1946   ($\alpha/3$) = $2\alpha/3$. (The weight w_{31} is for the arrow going from H_3 to H_1.)
1947   2. New alpha level at H_2 = old alpha level at H_2 + w_{32} × (the alpha level at H_3) = $\alpha/3$ + (0) ×
1948   ($\alpha/3$) = $\alpha/3$. (Note that there is no arrow shown from H_3 to H_2, as its weight $w_{32} = 0$.)
1949   3. New weight w_{12} for the arrow going from H_1 to H_2 = (old w_{12} + A)/(1- B), where
1950   A = additional weight for H_1 to H_2 going through H_3 = w_{13} × w_{32} = (1/3) × (0) = 0, and
1951   B = adjustment for the arrow going from H_1 to H_3 and returning back to H_1 = w_{13} × w_{31} =
1952   (1/3) × (1) = 1/3. Therefore, new w_{12} = (2/3 + 0) /(1 − 1/3) = 1.
1953   4. Similarly, new weight w_{21} for the arrow going from H_2 to H_1 = (old w_{21} + w_{23} × w_{31})/(1- w_{23}
1954   × w_{32}) = [1/2 + (1/2) × (1)]/ [1 - (1/2) × (0)] = 1.
1955
1956 This gives the updated graph in Figure A6 (b). Similar calculations can be made for graphs for
1957 H_1 and H_3 if H_2 is rejected and for H_2 and H_3 on rejecting H_1.
1958

(a)

(b)
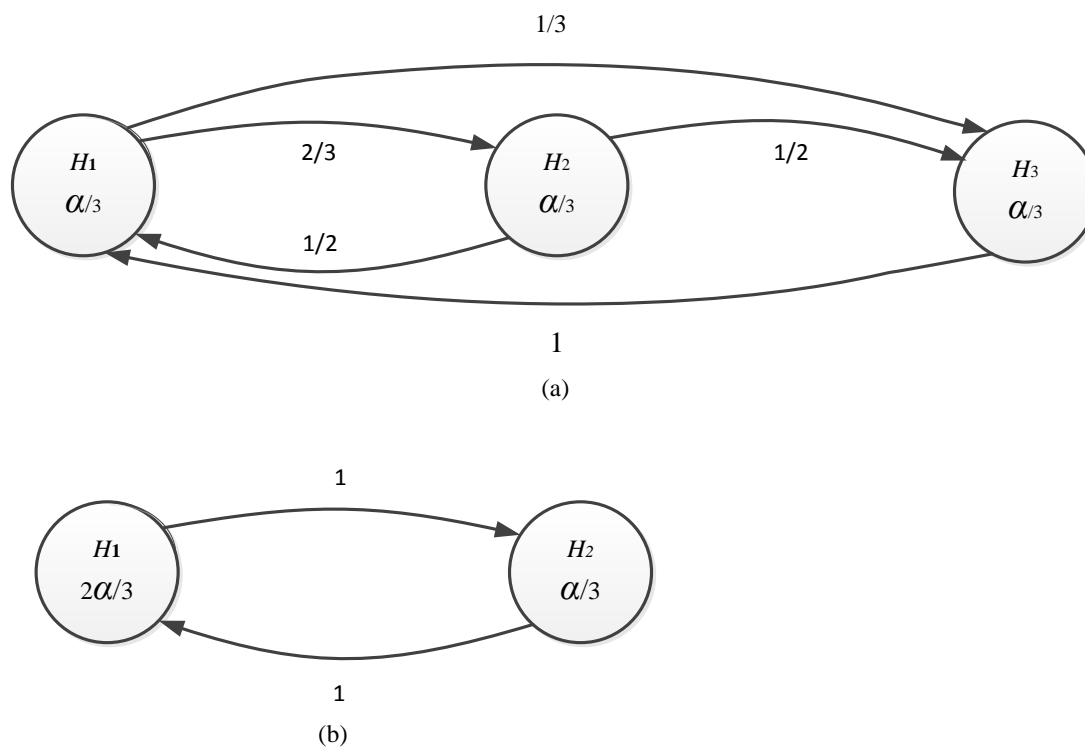
1959
1960 **Figure A6**: Initial diagram (a) for three hypotheses with fixed weights on the paths connecting
1961 the hypotheses, and updated graph (b) when hypotheses H_1 and H_2 are not yet rejected but H_3 is
1962 rejected.

1963