
Guidance for Industry

Non-Inferiority Clinical

Trials

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 90 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit 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 Robert Temple at 301-796-2270 or Robert O'Neill at 301-796-1700 (CDER), or the Office of Communication, Outreach, and Development (CBER) at 301-800-835-4709 or 301-827-1800.

**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)**

**March 2010
Clinical/Medical**

Guidance for Industry

Non-Inferiority Clinical

Trials

Additional copies are available from:

Office of Communication

Division of Drug Information, WO51, Room 2201

Center for Drug Evaluation and Research

Food and Drug Administration

10903 New Hampshire Ave.

Silver Spring, MD 20993

Phone: 301-796-3400; Fax: 301-847-8714

druginfo@fda.hhs.gov

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

or

Office of Communication, Outreach, and Development

Center for Biologics Evaluation and Research

Food and Drug Administration

1401 Rockville Pike, Rockville, MD 20852-1448

Phone: 800-835-4709 or 301-827-1800

ocd@fda.hhs.gov

<http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/default.htm>

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
March 2010
Clinical/Medical

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
II.	BACKGROUND	1
III.	GENERAL CONSIDERATION OF NON-INFERIORITY STUDIES: REGULATORY, STUDY DESIGN, SCIENTIFIC, AND STATISTICAL ISSUES.....	2
A.	Basic Principles of a Non-Inferiority Study	2
B.	Practical Considerations in Use of NI Designs.....	13
IV.	CHOOSING THE NON-INFERIORITY MARGIN AND ANALYZING THE RESULTS OF AN NI TRIAL	17
A.	Introduction	17
B.	Statistical Uncertainties in the NI Study and Quantification of Treatment Effect of Active Control.....	19
C.	Statistical Methods for NI Analysis	27
D.	Considerations for Selecting M_2, the Clinical Margin, and the Role of Subjective Judgment.....	31
E.	Estimating the Sample Size for an NI Study.....	32
F.	Potential Biases in an NI Study	33
G.	Role of Adaptive Designs in NI Studies — Sample Size Re-estimation to Increase the Size of an NI Trial	33
H.	Testing NI and Superiority in an NI Study	34
V.	COMMONLY ASKED QUESTIONS AND GENERAL GUIDANCE	35
	APPENDIX — EXAMPLES.....	40
	REFERENCES - EXAMPLES	56
	GENERAL REFERENCES.....	59

Contains Nonbinding Recommendations

Draft – Not for Implementation

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37

Guidance for Industry¹ Non-Inferiority Clinical Trials

This draft guidance, when finalized, will represent the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

I. INTRODUCTION

This guidance provides sponsors and review staff in the Center for Drug Evaluation and Research (CDER) and Center for Biologic Evaluation and Research (CBER) at the Food and Drug Administration (FDA) with our interpretation of the underlying principles involved in the use of non-inferiority (NI) study designs to provide evidence of the effectiveness of a drug or biologic.² The guidance gives advice on when NI studies can be interpretable, on how to choose the NI margin, and how to analyze the results.

II. BACKGROUND

This guidance consists of four parts. The first part is a general discussion of regulatory, study design, scientific, and statistical issues associated with the use of non-inferiority studies when these are used to establish the effectiveness of a new drug. The second part focuses on some of these issues in more detail, notably the quantitative analytical and statistical approaches used to determine the non-inferiority margin for use in NI studies, as well as the advantages and disadvantages of available methods. The third part addresses commonly asked questions about NI studies and provides practical advice about various approaches. The fourth part includes five examples of successful and unsuccessful efforts to define non-inferiority margins and conduct NI studies.³

FDA's guidance documents, including this guidance, do not establish legally enforceable responsibilities. Instead, guidance describes the Agency's current thinking on a subject and should be viewed as recommendations unless specific regulatory or statutory requirements

¹ This guidance has been prepared by the Office of Biostatistics and the Office of New Drugs in the Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER) at the Food and Drug Administration.

² For the purposes of this guidance, all references to *drugs* include both human drugs and therapeutic biologic products unless otherwise specified.

³ References: in this guidance, reference to methods or studies are not included in the text; rather they are included in a General Reference section and a separate reference section for the examples in the Appendix.

38 are cited. The use of the word *should* in Agency guidances means that something is
39 suggested or recommended, not that it is required.

40

41

42 **III. GENERAL CONSIDERATION OF NON-INFERIORITY STUDIES:**
43 **REGULATORY, STUDY DESIGN, SCIENTIFIC, AND STATISTICAL**
44 **ISSUES**

45

46 **A. Basic Principles of a Non-Inferiority Study**

47

48 *1. Superiority Trials versus Non-Inferiority Trials to Demonstrate Effectiveness*

49

50 FDA’s regulations on adequate and well-controlled studies (21 CFR 314.126) describe four
51 kinds of concurrently controlled trials that provide evidence of effectiveness. Three of them
52 — placebo, no treatment, and dose-response controlled trials — are superiority trials that
53 seek to show that a test drug is superior to the control (placebo, no treatment, or a lower dose
54 of the test drug). The fourth kind of concurrent control, comparison with an active treatment
55 (active control), can also be a superiority trial, if the intent is to show that the new drug is
56 more effective than the control. More commonly, however, the goal of such studies is to
57 show that the difference between the new and active control treatment is small, small enough
58 to allow the known effectiveness of the active control to support the conclusion that the new
59 test drug is also effective. How to design and interpret such studies so that they can support
60 such a conclusion is a formidable challenge.

61

62 These active control trials, which are not intended to show superiority of the test drug, but to
63 show that the new treatment is not inferior to an unacceptable extent, were once called
64 equivalence trials, but this is a misnomer, as true equivalence (i.e., assurance that the test
65 drug is not **any** less effective than the control), could only be shown by demonstrating
66 superiority. Because the intent of the trial is one-sided (i.e., to show that the new drug is not
67 materially worse than the control), they are now called non-inferiority (NI) trials. But that
68 too, is a misnomer, as guaranteeing that the test drug is not any (even a little) less effective
69 than the control can only be demonstrated by showing that the test drug is superior. What
70 non-inferiority trials seek to show is that any difference between the two treatments is small
71 enough to allow a conclusion that the new drug has at least some effect or, in many cases, an
72 effect that is not too much smaller than the active control.

73

74 The critical difference between superiority and NI trials is that a properly designed and
75 conducted superiority trial, if successful in showing a difference, is entirely interpretable
76 without further assumptions (other than lack of bias or poor study conduct); that is, the result
77 speaks for itself and requires no further extra-study information. In contrast, the NI study is
78 dependent on knowing something that is not measured in the study, namely, that the active
79 control had its expected effect in the NI study. This is critical to knowing that the trial had
80 *assay sensitivity* (i.e., could have distinguished an effective from an ineffective drug). A
81 successful superiority trial has, by definition, assay sensitivity. A “successful” NI trial, one
82 that shows what appears to be an acceptably small difference between treatments, may or

Contains Nonbinding Recommendations

Draft – Not for Implementation

83 may not have had assay sensitivity and may or may not have supported a conclusion that the
84 test drug was effective. Thus, if the active control had no effect at all in the NI trial (i.e., did
85 not have any of its expected effect), then finding even a very small difference between
86 control and test drug is meaningless, providing no evidence that the test drug is effective.
87 Knowing whether the trial had assay sensitivity relies heavily on external (not within-study)
88 information, giving NI studies some of the characteristics of a historical control trial.

89
90 FDA regulations have recognized since 1985 the critical need to know, for an NI trial to be
91 interpretable, that the active control had its expected effect in the trial. Thus, 21 CFR
92 314.126(a)(2)(iv), unchanged since 1985, says:

93
94 If the intent of the trial is to show similarity of the test and control drugs, the report of
95 the study should assess the ability of the study to have detected a difference between
96 treatments. Similarity of test drug and active control can mean either that both drugs
97 were effective or that neither was effective. The analysis of the study should explain
98 why the drugs should be considered effective in the study, for example, by reference to
99 results in previous placebo-controlled studies of the active control drug.

100

101 2. *Logic of the NI Trial*

102

103 In a placebo-controlled trial, the null hypothesis (H_0) is that the response to the test drug (T)
104 is less than or equal to the response to the placebo (P); the alternative hypothesis (H_a) is that
105 the response to the test drug is greater than P.

106

$$107 H_0: T \leq P; \quad T - P \leq 0$$

$$108 H_a: T > P; \quad T - P > 0$$

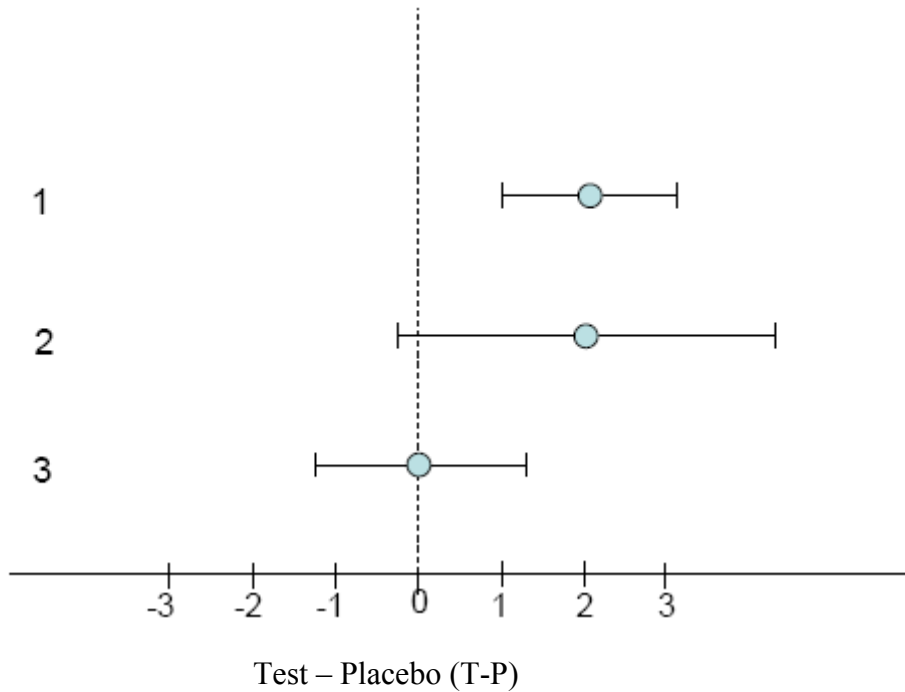
109

110 In most cases, a treatment effect is established statistically by showing that the lower bound
111 of the two-sided 95% confidence interval (equivalent to the lower bound of a one-sided
112 97.5% confidence interval) for T-P is > 0 .⁴ This shows that the effect of the test drug is
113 greater than 0. See Figure 1.

⁴ Ref. 4

114
115
116

Figure 1: Three Possible Results of a Placebo-Controlled Superiority Study (Point Estimate, 95% CI)



117
118
119
120
121
122
123
124
125
126
127

1. Point estimate of effect is 2; 95% CI lower bound is 1. Conclusion: Drug is effective and appears to have an effect of at least 1.
2. Point estimate of effect is 2; 95% CI lower bound is <0 (study perhaps too small). Conclusion: Drug is not shown to be effective.
3. Point estimate of effect is 0; 95% CI lower bound is well below 0. Conclusion: Drug shows no suggestion of effectiveness.

128
129
130
131
132
133
134
135

In an NI study whose goal is to show that the new drug has an effect greater than zero, the null hypothesis is that the degree of inferiority of the new drug (T) to the control (C), C-T, is greater than the non-inferiority margin M_1 , where M_1 represents what is thought to be the whole effect of the active control (C) relative to placebo in the NI study.⁵

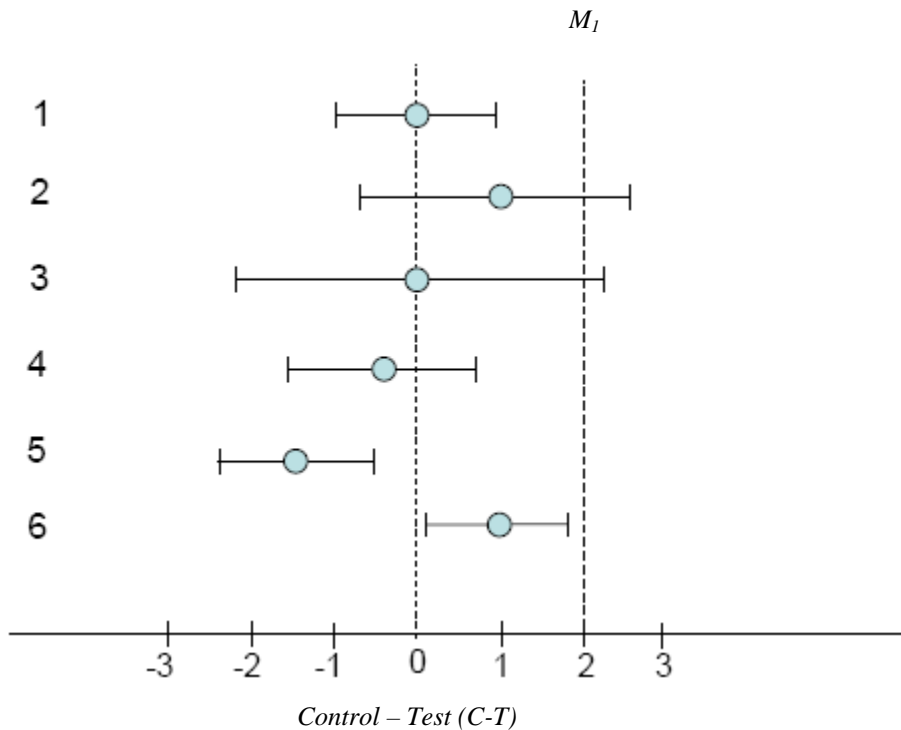
$$H_0: C - T \geq M_1 \text{ (T is inferior to the control by } M_1 \text{ or more)}$$

$$H_a: C - T < M_1 \text{ (T is inferior to the control by less than } M_1)$$

⁵ M is the non-inferiority margin used in the NI study. It can be no larger than the entire effect that C is presumed to have had in the study, in which case it is called M_1 . As described below, the margin of interest can be smaller than M_1 , in which case it is called M_2 .

136 Again, non-inferiority is established by showing that the upper bound of the two-sided
 137 confidence interval for C-T is $< M_1$. If the chosen M_1 does in fact represent the entire effect
 138 of the active control drug in the NI study, a finding of non-inferiority means that the test drug
 139 has an effect greater than 0 (see Figure 2). Thus, in the non-inferiority setting, assay
 140 sensitivity means that the control drug had at least the effect it was expected to have (i.e.,
 141 M_1).

142
 143 **Figure 2: Results of NI Study Showing C-T and 95% CI**
 144 **($M_1 = 2$)**



- 150 1. Point estimate of C-T is 0, suggesting equal effect; upper bound of the 95% CI for C-T
 151 is 1, well below M_1 ; NI is demonstrated.
- 152 2. Point estimate of C-T favors C; upper bound of the 95% CI for C-T is >2 , well above
 153 M_1 ; NI is not demonstrated.
- 154 3. Point estimate of C-T is zero, suggesting equal effect; but upper bound of the 95% CI
 155 for C-T is >2 (i.e., above M_1), so that NI is not demonstrated.
- 156 4. Point estimate favors T; NI is demonstrated, but superiority is not demonstrated.
- 157 5. Point estimate favors T; superiority and NI are demonstrated.
- 158 6. Point estimate of C-T favors C and C is statistically significantly superior to T.
 159 Nonetheless, upper bound of the 95% CI for C-T < 2 (M_1), so that NI is also
 160 demonstrated for the NI margin M_1 . (This outcome would be unusual and could
 161 present interpretive problems.)
 162

163 The critical problem, and the major focus of this guidance, is determining M_1 , which is not
164 measured in the NI study (there is no concurrent placebo group). It must be estimated (really
165 assumed) based on the past performance of the active control and by comparison of prior test
166 conditions to the current test environment (see section III.A.4). Determining the NI margin
167 is the single greatest challenge in the design, conduct, and interpretation of NI trials.

168
169 The choice of the margin M_1 has important practical consequences. The smaller the margin,
170 the smaller the upper bound of the 95% two-sided confidence interval for C-T must be, and
171 the larger the sample size that will be needed.

172

173 3. Reasons for Using a Non-Inferiority Design

174

175 The usual reason for using a non-inferiority active control study design instead of a study
176 design having more readily interpretable results (i.e., a superiority trial) is an ethical one.
177 Specifically, this design is chosen when it would not be ethical to use a placebo, or a no-
178 treatment control, or a very low dose of an active drug, because there is an effective
179 treatment that provides an important benefit (e.g., life-saving or preventing irreversible
180 injury) available to patients for the condition to be studied in the trial. Whether a placebo
181 control can be used depends on the nature of the benefits provided by available therapy. The
182 International Conference on Harmonization guidance E10 on *Choice of Control Group and*
183 *Related Issues in Clinical Trials* (ICH E10) states:

184

185 In cases where an available treatment is known to prevent serious harm, such as death
186 or irreversible morbidity in the study population, it is generally inappropriate to use a
187 placebo control. [The term “generally” leaves room for a placebo control if the
188 known effective treatment is very toxic.]

189

190 In other situations, where there is no serious harm, it is generally considered ethical
191 to ask patients to participate in a placebo-controlled trial, even if they may experience
192 discomfort as a result, provided the setting is non-coercive and patients are fully
193 informed about available therapies and the consequences of delaying treatment.

194

195 There are, however, other reasons for using an active control: (1) interest in comparative
196 effectiveness and (2) assessing the adequacy (assay sensitivity) of a placebo-controlled study.
197 These are not the focus of this guidance, but will be considered briefly.

198

199 a. Comparative effectiveness

200

201 There is growing interest among third party payers and some regulatory authorities, on both
202 cost effectiveness and medical grounds, in the comparative effectiveness of treatments, and
203 an increasing number of such studies are being conducted. A critical issue is the importance
204 of including a placebo group, as well as the active comparator, in such studies (a 3-arm trial)
205 to assess assay sensitivity (i.e., the ability of the trial to detect differences of a specified size
206 between treatments). When the treatment is clinically critical, it will, of course, not be
207 ethically acceptable to include a placebo group, and the discussion of NI studies that follows
208 will be highly relevant to such trials. Even where it would be ethical to include a placebo

Contains Nonbinding Recommendations

Draft – Not for Implementation

209 group in addition to the active treatments (e.g., in studies of a symptomatic treatment), one is
210 not necessarily included in these comparative trials. Such omission of a placebo group may
211 render such studies uninformative, however, when they show no difference between
212 treatments, unless assay sensitivity can be supported in some other way.

213

214 Where comparative effectiveness is the principal interest, it is usually important—where it is
215 ethical, as would be the case in most symptomatic conditions—to include a placebo control
216 as well as the active control. Trials of most symptomatic treatments have a significant failure
217 rate (i.e., they often cannot show the drug is superior to placebo). Where that is the case in a
218 comparative trial, seeing no difference between treatments is uninformative. Inclusion of a
219 placebo group can provide clear evidence that the study did have assay sensitivity (the ability
220 to distinguish effective from ineffective treatments), critical if a finding of no difference
221 between treatments is to be interpretable. For example, we have seen that approximately
222 50% of all placebo-controlled antidepressant trials of effective agents cannot distinguish drug
223 from placebo. A trial in which two antidepressants are compared and found to have a similar
224 effect is informative only if we know that the two drugs can be distinguished from the
225 concurrent placebo group.

226

227 b. Assessing assay sensitivity of a placebo-controlled study

228

229 Although a successful superiority trial (e.g., placebo-controlled) is readily interpreted, a
230 failed trial of this design is not. Failure to show superiority to placebo can mean that the
231 drug is ineffective or that the trial lacked assay sensitivity. To distinguish between these two
232 possibilities, it is often useful to include an active control in placebo-controlled studies of
233 drugs in a class or condition where known effective drugs often cannot be distinguished from
234 placebo (e.g., depression, allergic rhinitis, angina, and many other symptomatic conditions).
235 If the active control is superior to placebo but the test drug is not, one can conclude that the
236 test drug lacks effectiveness (or at least is less effective than the active control). If neither
237 the active control nor the test drug is superior to placebo, the trial lacked assay sensitivity and
238 is uninformative about the effect of the test drug.

239

240 4. *The Non-Inferiority Margin*

241

242 As described above, the NI study seeks to show that the difference in response between the
243 active control (C) and the test drug (T), (C-T), the amount by which the control is superior to
244 test drug, is less than some pre-specified non-inferiority margin (M). M can be no larger than
245 the presumed entire effect of the active control in the NI study, and the margin based on that
246 whole active control effect is generally referred to as M_1 . It is critical to reiterate that M_1 is
247 not measured in the NI trial, but must be assumed based on past performance of the active
248 control, the comparison of the current NI study with prior studies, and assessment of the
249 quality of the NI study (see below). The validity of any conclusion from the NI study
250 depends on the choice of M_1 . If, for example, the NI margin is chosen as 10 (because we are
251 sure the control had an effect of at least that size), and the study does indeed rule out a
252 difference of 10 (seeming to demonstrate “effectiveness” of T), but the true effect of C in this
253 study was actually less than 10, say 5, T would not in fact have been shown to have any

Contains Nonbinding Recommendations

Draft – Not for Implementation

254 effect at all; it will only appear to have had such an effect. The choice of M_1 , and assurance
255 that this effect was present in the trial (i.e., the presence of assay sensitivity) is thus critical to
256 obtaining a meaningful, correct answer in an NI study.

257

258 Because the consequence of choosing a margin greater than the actual treatment effect of the
259 active control in the study is the false conclusion that a new drug is effective (a very bad
260 public health outcome), there is a powerful tendency to be conservative in the choice of
261 margin and in the statistical analysis that seeks to rule out a degree of inferiority of the test
262 drug to the active control of more than that margin. This is generally done by ensuring that
263 the upper bound of the 95% two-sided confidence interval for C-T is smaller than M_1 . The
264 upper bound of the confidence interval for C-T is not, however, the only measurement of
265 interest, just as the lower bound of a 95% confidence interval for effect size of drug versus
266 placebo is not the only value of relevance in a placebo-controlled trial. The point estimate of
267 the treatment effect and the distribution of estimates of C-T smaller than the 95% upper
268 bound are also relevant. Nonetheless, the upper bound of the 95% CI is typically used to
269 judge the effectiveness of the test drug in the NI study, just as a two-sided p-value of 0.05 or
270 less is traditionally the standard used for defining success in a superiority trial. The 95% CI
271 upper bound for C-T is used to provide a reasonably high level of assurance that the test drug
272 does, in fact, have an effect greater than zero (i.e., that it has not lost all of the effect of the
273 active control).

274

275 Although the NI margin used in a trial can be no larger than the entire assumed effect of the
276 active control in the NI study (M_1), it is usual and generally desirable to choose a smaller
277 value, called M_2 , for the NI margin. Showing non-inferiority to M_1 would provide assurance
278 that the test drug had an effect greater than zero. However, in many cases that would not be
279 sufficient assurance that the test drug had a clinically meaningful effect. After all, the reason
280 for using the NI design is the perceived value of the active control drug. It would not usually
281 be acceptable to lose most of that active control's effect in a new drug. It is therefore usual
282 in NI studies to choose a smaller margin (M_2) that reflects the largest loss of effect that
283 would be clinically acceptable. This can be described as an absolute difference in effect
284 (typical of antibiotic trials) or as a fraction of the risk reduction provided by the control
285 (typical in cardiovascular outcome trials). Note that the clinically acceptable margin could
286 be relaxed if the test drug were shown to have some important advantage (e.g., on safety or
287 on a secondary endpoint).

288

289 The definitions used to describe these two versions of M are:

290

291 M_1 = the entire effect of the active control assumed to be present in the NI study
292 M_2 = the largest clinically acceptable difference (degree of inferiority) of the test drug
293 compared to the active control

294

295 M_1 is based on (1) the treatment effect estimated from the historical experience with the
296 active control drug, (2) assessment of the likelihood that the current effect of the active
297 control is similar to the past effect (the constancy assumption), and (3) assessment of the
298 quality of the NI trial, particularly looking for defects that could reduce a difference between

299 the active control and the new drug (this diminution of the between-treatment difference is a
 300 “bias toward the null” in a trial seeking to show a difference (i.e., superiority), but in this
 301 case is a “bias toward the alternative”). Note that because of this third element, the size of
 302 M_1 cannot be entirely specified until the NI study is complete.

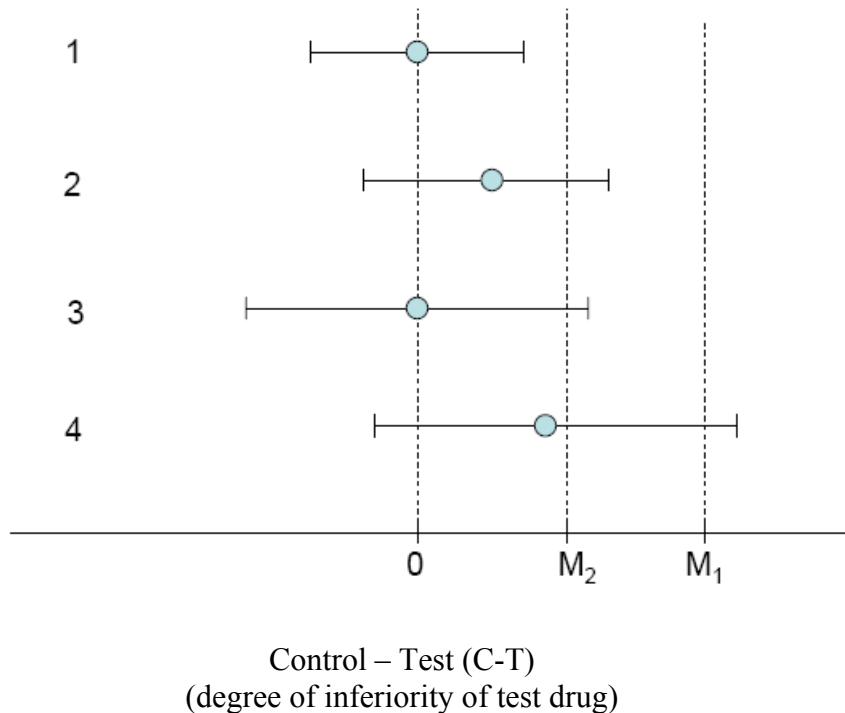
303

304 M_2 is a matter of clinical judgment, but M_2 can never be greater than M_1 , even if, for active
 305 control drugs with small effects, a clinical judgment might argue that a larger difference is
 306 not clinically important. Even if that clinical judgment were reasonable, an M_2 greater than
 307 M_1 cannot be used to demonstrate that the test drug has any effect. As explained above,
 308 ruling out a difference between the active control and test drug larger than M_1 is the critical
 309 finding that supports a conclusion of effectiveness. This analysis is approached with great
 310 rigor; that is, a difference (C-T) larger than M_1 needs to be ruled out with a high degree of
 311 statistical assurance. As M_2 represents a clinical judgment, there may be a greater flexibility
 312 in interpreting a 95% upper bound for C-T that is slightly greater than M_2 , as long as the
 313 upper bound is still well less than M_1 (see Figure 3).

314

315 **Figure 3. Active Control – Test Drug differences**
 316 **(Point estimate, 95% CI)**

317



318

319

320

321

322

323

324

325

326

327

1. C-T point estimate = 0 and upper bound of 95% CI < M_2 , indicating test drug is effective (NI demonstrated).
2. Point estimate of C-T favors C and upper bound of 95% CI < M_1 but > M_2 , indicating effect > 0 but unacceptable loss of the control effect.
3. Point estimate of C-T is zero and upper bound of 95% CI < M_1 but it is

Contains Nonbinding Recommendations

Draft – Not for Implementation

328 slightly greater than M_2 . Judgment could lead to conclusion of effectiveness.
329 4. C-T point estimate favors C and upper bound of 95% CI $> M_1$, indicating
330 there is no evidence of effectiveness for test drug.

331

332

333

5. Assay Sensitivity and Choosing M_1

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

- **Historical evidence of sensitivity to drug effects (HESDE) (ICH E-10)**

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

HESDE means that appropriately designed and conducted trials in the past that used a specific active treatment (generally the one that is to be used in the new NI study or, in some cases, one or more pharmacologically closely related drugs) regularly showed this treatment to be superior to placebo (or some other treatment). These consistent findings allow for a reliable estimate of the drug's effect size compared to placebo in those past studies, a reasonable starting point for estimating its effect in the NI study. The estimate of effect size must take the variability of past results into account; one would not presume that the largest effect seen in any trial, or even the point estimate of a meta-analysis, is likely to be the effect size in the new study. Analysis of historical data will be discussed further in section IV.

HESDE cannot be determined for many symptomatic treatments, where well-designed and conducted studies often fail to distinguish drug from placebo (e.g., treatments for depression, anxiety, insomnia, angina, symptomatic heart failure, symptoms of irritable bowel disease, and pain). In those cases, there is no reason to assume that an active control would have shown superiority to a placebo (had there been one) in any given NI study, and NI studies of drugs for these treatments are not informative. This is also true for some outcome effectiveness findings, such as secondary prevention of cardiovascular disease with aspirin and post-infarction beta blockade. In the case of aspirin, the largest placebo-controlled trial

373 (AMIS, the Aspirin Myocardial Infarction Study; see Example 3) showed no effect of aspirin
374 at all, even though other trials all favored aspirin. Similarly, of more than 30 post-infarction
375 beta-blocker trials, only a small number showed significantly improved survival or other
376 cardiovascular benefit.

377

- 378 • **Similarity of the current NI trial to the historical studies – the “constancy**
379 **assumption”**

380

381 The conclusion that HESDE can be used to choose M_1 for the new NI study can be reached
382 only when it is possible to conclude that the NI study is sufficiently similar to the past studies
383 with respect to all important study design and conduct features that might influence the effect
384 size of the active control. This is referred to as the “constancy assumption.” The design
385 features of interest include the characteristics of the patient population, important
386 concomitant treatments, definitions and ascertainment of study endpoints, dose of active
387 control, entry criteria, and analytic approaches. The effect of an ACE inhibitor on heart
388 failure mortality has repeatedly been shown in studies where the drugs were added to
389 diuretics and digoxin, but evolution in treatment since those studies were conducted raises
390 questions about our understanding of the present-day effect of these drugs. Since the time of
391 those studies, new medications (beta blockers, spironolactone) have come into standard use.
392 We do not know whether the past effect would still be present when ACE inhibitors are
393 added to a regimen including those two drugs. Similarly, the effect of a thrombolytic on
394 cardiovascular mortality could depend on how soon after symptoms the drug was given,
395 concomitant use of anticoagulants and platelet inhibitors, and use of lipid-lowering drugs.
396 As a general matter, the historical and new NI studies should be as close to identical as
397 possible in all important respects.

398

399 It is easier to be reasonably assured that endpoints in the historical trial will be similar to, and
400 will be evaluated similarly to, endpoints in the new trial when these are well-standardized
401 and objective. The effect of the active control could be on a single endpoint (e.g., mortality)
402 or on a composite (e.g., death, heart attack, and stroke), but, again, it is critical that
403 measurement and assessment of these be reasonably consistent over time. The endpoint used
404 in the NI study need not necessarily be the one used in the original trials of the active control
405 if data are available to estimate the occurrence rate of the new endpoint used in the NI study.
406 For example, even if the historical studies used a mortality endpoint, the studies could be
407 used if data could be obtained to calculate an effect size for death plus hospitalization, so
408 long as it was possible to be confident that the circumstances leading to the hospitalization
409 were similar in the historical studies and the NI study. Note, however, that it would not be
410 acceptable to search through a range of endpoints to find the largest historical effect, as this
411 could represent an overestimate of the effect to be expected in the NI study.

412

413 In general, where there has been substantial evolution over time in disease definition and
414 treatment, supporting the constancy assumption may be difficult.

415

416 Although an NI study can be designed to be similar in most aspects to the historical studies, it
417 may not be possible to assess that similarity fully until the NI study is completed and various

418 characteristics of the study population and response are evaluated. When there is known
419 demonstrated heterogeneity of the active control treatment effect related to patient
420 characteristics (e.g., age, gender, severity), and when that heterogeneity can be quantified, it
421 may be necessary to adjust the estimate of the active control effect size in the NI study if the
422 mix of patient characteristics in the historical and NI studies differ substantially.

423

424 The property of constancy of the treatment effect may depend on which metric is chosen to
425 represent the treatment effect. This issue is discussed in more depth in section IV.B.2.d.
426 Experience suggests that when background rates of outcomes differ among study
427 populations, metrics like hazard ratios or relative risks are more stable than is a metric like
428 absolute effect size, which is more sensitive to changes in event rates in the population.

429

430 **• Good Study Quality**

431

432 A variety of study quality deficiencies can introduce what is known as a “bias toward the
433 null,” where the observed treatment difference in an NI study is decreased from the true
434 difference between treatments. These deficiencies include imprecise or poorly implemented
435 entry criteria, poor compliance, and use of concomitant treatments whose effects may overlap
436 with the drugs under study, inadequate measurement techniques, or errors in delivering
437 assigned treatments. Many such defects have small (or no) effects on the variability of
438 outcomes (variance) but reduce the observed difference C-T, potentially leading to a false
439 conclusion of non-inferiority. It should also be appreciated that intent-to-treat approaches,
440 which preserve the principle that all patients are analyzed according to the treatment to which
441 they have been randomized even if they do not receive it, although conservative in
442 superiority trials, are not conservative in an NI study, and can contribute to this bias toward
443 the null. It is more important than usual to plan in advance steps to ensure quality during the
444 conduct of an NI study.

445

446 Finally, it should be recognized that although most investigators seek to carry out high
447 quality trials, the incentives in an NI study are perverse, and quite different from those in
448 superiority trials. In a superiority trial, sloppiness can lead to study failure, and major efforts
449 in trial conduct and monitoring are therefore devoted to avoiding it. In general, sloppiness of
450 any sort obscures true treatment differences. In an NI trial, in contrast, where the goal is to
451 show no difference (or no difference greater than M), poor quality can sometimes lead to an
452 apparent finding of non-inferiority that is incorrect. There is therefore a critical need for
453 particular attention to study quality and conduct when planning and executing an NI study.

454

455 *6. Regulatory Conclusions*

456

457 A successful non-inferiority study shows rigorously that the test drug has an effect greater
458 than zero if it excludes an NI margin of M_1 , so long as M_1 is well chosen and represents an
459 effect that the control drug actually would have had (versus a placebo, had there been a
460 placebo group). It can also show that the test drug had an effect greater than some fraction of
461 the control drug effect, depending on the M_2 that is used. It should be appreciated that in
462 addition to the rigorous demonstration of effectiveness, the trial provides additional

463 information, just as a placebo-controlled trial supporting the effectiveness of a drug does.
464 The point estimate of the drug effect and its confidence interval (usually 95% but could be
465 90% or 99% under some circumstances) provides information about how large the difference
466 in treatment effect between the test and control drug is likely to be.

467
468 In most cases a successful NI study supports effectiveness of the test drug, but it only rarely
469 will support a conclusion that the drug is “equivalent” or “similar” to the active control, a
470 concept that has not been well-defined for these situations. Such similarity might be
471 concluded, however, if the point estimate of the test drug favored it over the control and the
472 upper bound of the 95% CI for C-T was close to showing superiority. Where the chosen M_2
473 is very small compared to the control drug effect (e.g., a 10% margin in an antibiotic trial in
474 urinary tract infections where response rate is 80%), it might be concluded that the
475 effectiveness of the test drug and control are very similar.

476

477 **B. Practical Considerations in Use of NI Designs**

478

479 *1. Consider Alternative Designs*

480

481 ICH E10 identifies a wide variety of study designs that may be better than an NI design in
482 situations where there is difficulty or uncertainty in setting the NI margin, or where the NI
483 margin needs to be so small that the NI study sample size becomes impossibly large.

484

485 • **Add-on study**

486

487 In many cases, for a pharmacologically novel treatment, the most interesting question
488 is not whether it is effective alone but whether the new drug can add to the
489 effectiveness of treatments that are already available. The most pertinent study would
490 therefore be a comparison of the new agent and placebo, each added to established
491 therapy. Thus, new treatments for heart failure have added new agents (e.g., ACE
492 inhibitors, beta blockers, and spironolactone) to diuretics and digoxin. As each new
493 agent became established, it became part of the background therapy to which any new
494 agent and placebo would be added. This approach is also typical in oncology, in the
495 treatment of seizure disorders, and, in many cases, in the treatment of AIDS.

496

497 • **Identifying a population not known to benefit from available therapy in which a 498 placebo-controlled trial is acceptable**

499

500 In many outcome study settings, effectiveness is established for some clinical settings
501 (e.g., severe disease) but not others. Therefore, it may be possible to study less
502 severely ill patients in placebo-controlled trials. The demonstration that simvastatin
503 was effective in hypercholesterolemic post-infarction patients (4S), for example, did
504 not forestall studies of statins in hypercholesterolemic non-infarction patients
505 (WOSCOPS) or in patients with lesser degrees of hypercholesterolemia (TEXCAPS).
506 This is legitimate so long as one does not in fact know the treatment is of value in the
507 new study population. Recently, it has been possible to study angiotensin receptor

508 blockers (ARBs) in heart failure in a placebo-controlled trial in patients intolerant of
509 ACE inhibitors (known to improve survival). It would not have been possible to deny
510 a more general population of heart failure patients an ACE inhibitor.

511

512 • **Early escape, rescue treatment, randomized withdrawal**

513

514 In symptomatic conditions, there may be reluctance to leave people on placebo for
515 prolonged periods when effective therapy exists. It is possible to incorporate early
516 escape/rescue provisions for patients who do not respond by a particular time, or to
517 use a design that terminates patients on first recurrence of a symptom such as unstable
518 angina, grand mal seizure, or paroxysmal supra-ventricular tachycardia. To evaluate
519 the persistence of effects over time, where conducting a long-term placebo-controlled
520 trial would be difficult, a randomized withdrawal study can be used. Such a study
521 randomly assigns patients treated with a drug for a long period to placebo or
522 continued drug treatment. As soon as symptoms return, the patient is considered to
523 have had an endpoint. This design was first suggested to evaluate long-term benefit
524 in angina.

525

526 2. *Number of Studies Needed*

527

528 Ordinarily, with exceptions allowed by the FDA Modernization Act of 1997 (the
529 Modernization Act), FDA expects that there will be more than one adequate and well-
530 controlled study supporting effectiveness. The Modernization Act allows one study plus
531 confirmatory evidence to serve as substantial evidence in some cases, and FDA has discussed
532 in guidance (*Providing Clinical Evidence of Effectiveness for Human Drug and Biological*
533 *Products*) when a single study might be sufficient.

534

535 Where there is uncertainty about the historical effect size (and thus M_1) because of variability
536 or reliance on a single historical study, it will usually be necessary to have more than one NI
537 study to support effectiveness.

538

539 Where the studies are of relatively modest size (e.g., most antibiotic NI trials), there is no
540 impediment to conducting more than one NI trial. When the trials needed are very large (to
541 have adequate statistical power), however, this may become a significant problem and it is
542 worth considering what might make a single trial persuasive. Generally, two considerations
543 might do so: (1) prior information, (2) a statistically persuasive result.

544

545 • **Prior information**

546

547 It is common in NI trials for the test drug to be pharmacologically similar to the active
548 control. (If they were not pharmacologically similar, an add-on study would usually have
549 been more persuasive and more practical). In that case, the expectation of similar
550 performance (but still requiring confirmation in a trial) might make it possible to accept a
551 single trial and perhaps could also allow less conservative choices in choosing the non-

552 inferiority margin. A similar conclusion might be reached when other types of data are
553 available, for example:

- 554 • If there were a very persuasive biomarker confirming similar activity of the test drug
555 and active control (e.g., tumor response, ACE inhibition, or extent of beta blockage)
- 556 • If the drug has been shown to be effective in closely-related clinical settings (e.g.,
557 effective as adjunctive therapy with an NI study of monotherapy)
- 558 • If the drug has been shown to be effective in distinct but related populations (e.g.,
559 pediatric versus adult)

560

- 561 • **Statistically persuasive result**

562

563 A conclusion that an NI trial can be considered statistically persuasive can be reached in
564 several ways, including the internal consistency of the NI finding, and the margin that is
565 ruled out with a two-sided 95% confidence interval. It is important to recognize that there
566 are two margins of interest, M_1 and M_2 . In an NI study, the clinically determined margin M_2
567 is smaller, often considerably smaller, than M_1 , which addresses the question of whether the
568 test drug has any effect. For example, M_2 might be chosen to be 40% of M_1 . By meeting
569 this M_2 criterion, ruling out a loss of 40% of the effect of the control, a single NI study
570 provides reasonable assurance that the test drug preserves a clinically sufficient fraction (at
571 least 60%) of the effect of the control treatment. At the same time, it provides strong
572 assurance (probably equivalent in strength to $p \leq 0.001$ in a superiority trial) that the test drug
573 has an effect greater than zero. Particularly where there is strong prior information on the
574 effectiveness of the pharmacological class being studied in the NI trial, showing non-
575 inferiority using M_2 thus provides very strong evidence, analogous statistically to the 2
576 studies (at $p \leq 0.05$) standard for difference–showing trials, that the new drug has an effect.
577 In such cases, a single such trial would usually be a sufficient basis for approval. Where the
578 effect of the drug is particularly critical, of course, it might be considered necessary to
579 demonstrate that loss of M_2 has been ruled out in more than one study.

580

581 In some cases, a study planned as an NI study may show superiority to the active control.
582 ICH E-9 and FDA policy has been that such a superiority finding arising in an NI study can
583 be interpreted without adjustment for multiplicity. Showing superiority to an active control
584 is very persuasive with respect to the effectiveness of the test drug, because demonstrating
585 superiority to an active drug is much more difficult than showing superiority to placebo.
586 Similarly, a finding of less than superiority, but with a 95% CI upper bound for C-T
587 considerably smaller than M_2 , is also statistically persuasive.

588

589 3. *Statistical Inferences*

590

591 The designer of an NI trial might hope that the test drug is actually superior to the control. It
592 is possible to design the NI study to first test the hypothesis of NI with the pre-specified
593 margin, and then if this test is successful, proceed to analyze the study for a superiority
594 conclusion. This sequential strategy is entirely acceptable. No statistical adjustment is
595 required. A possibility that has thus far had relatively little attention is to have different
596 endpoints with different goals (e.g., superiority on the composite endpoint of death, AMI,

597 and stroke, but NI on death alone). The multiple endpoints would require some alpha
598 adjustment in such a case, but the procedures here are not well defined. Similarly, if a study
599 had several doses, with interest in NI on each of them and, at the same time, interest in a
600 potential superiority finding for one or more doses, the analytical approach is not yet fully
601 established, although it is clear that some correction for multiplicity would be needed.

602

603 Seeking an NI conclusion in the event of a failed superiority test would almost never be
604 acceptable. It would be very difficult to make a persuasive case for an NI margin based on
605 data analyzed with study results in hand. If it is clear that an NI conclusion is a possibility,
606 the study should be designed as an NI study.

607

608 *4. Choice of Active Control*

609

610 The active control must be a drug whose effect is well-defined. The most obvious choice is
611 the drug used in the historical placebo-controlled trials. Where studies of several
612 pharmacologically similar drugs have been pooled, which is often done to obtain a better
613 estimate of effect and a narrower confidence interval, and thus a larger M_1 , the choice may
614 become complicated. In general, if the drugs in a meta-analysis of placebo-controlled trials
615 seem to have similar effects, any of them could be used as an active control. If their
616 observed treatment effects differ, however, even if not significantly, the one with the highest
617 point estimate of effect should ordinarily be used.

618

619 *5. Choice of NI Method*

620

621 The various approaches to calculating the NI margin and analyzing an NI study will be
622 discussed in detail in section IV, but the most straightforward and most readily understood
623 approach will be described here. This method is generally referred to as a fixed margin
624 method and the 95%-95% method (or 90%-95% method, depending on the CIs used to
625 calculate the NI margin) method. The first 95% refers to the confidence interval used to
626 choose the effect size from the historical data, and the second 95% refers to the confidence
627 level used to reject the null hypothesis in the NI study. This approach is illustrated by FDA's
628 evaluation of thrombolytics (TPA). To calculate the NI margin, all available placebo-
629 controlled trials of streptokinase, the active comparator or control, were pooled, giving a
630 point estimate for the effect on survival of a 25% reduction in mortality, with a one-sided
631 95% lower bound of 22%. As 22% represented the risk reduction by streptokinase compared
632 to placebo, this was translated to the risk increase from being on placebo ($1 \div .78$, or 1.28).
633 The NI study would therefore have had to rule out a 28% increase in risk (the risk increase
634 from a placebo) from not being on TPA. There was a clinical decision to ensure that not
635 more than 50% of the effect of streptokinase was lost, giving an NI margin (M_2) of 1.14, the
636 95% upper bound of the relative risk for TPA versus streptokinase (see section IV.B.2.c for
637 further discussion of this calculation).

638

639 This approach is relatively conservative, as it keeps separate the variability of estimates of
640 the treatment effect in the historical studies and the variability observed in the NI study, and
641 uses a fixed value for the estimate of the control effect based on historical data (the 90% or

642 95% CI lower bound), a relatively conservative estimate of the control drug effect. On the
643 other hand, a conservative estimate of an important endpoint such as mortality is not
644 necessarily unreasonable, particularly given the uncertainties associated with an NI design.

645

646 **IV. CHOOSING THE NON-INFERIORITY MARGIN AND ANALYZING THE** 647 **RESULTS OF AN NI TRIAL**

648

649 **A. Introduction**

650

651 This section will discuss how to determine the magnitude of the largest acceptable non-
652 inferiority margin, M_1 , and the clinical margin, M_2 , and how to analyze the NI study. M_1 is
653 the effect the active control (also called positive control) is thought to have had in the NI
654 study. As the effect of the active control in the NI study is not measured (there is no placebo
655 group), this effect must be assumed. The assumed value is based on the analysis of the effect
656 of the active control seen in past controlled studies. M_2 reflects the clinical judgment about
657 how much of M_1 should be preserved by ruling out a loss of M_2 . Thus, if it were concluded
658 that it would be necessary for a test drug to preserve 75% of a mortality effect, M_2 would be
659 25% of M_1 , the loss of effect that must be ruled out. It must be appreciated that subjectivity
660 and judgment are involved in all aspects of these determinations, a fundamental difference
661 from a superiority study where all the critical information is measured and no assumptions
662 are needed. This guidance will address how these judgments should be made in selecting the
663 margin selection specified in the NI analysis.

664

665 As described in section III, the selection of a margin for an NI study is a two-step process.
666 The first step involves making a reasonable assumption about the effect of the active
667 comparator in the NI study. M_1 is chosen to equal that treatment effect. If the advantage of
668 the control over the test drug in the NI study is larger than M_1 , then the test drug has not been
669 shown to have any effect. Effectiveness is therefore demonstrated by showing that the
670 advantage of the control over the test drug (C-T) is smaller than M_1 . This can be
671 demonstrated by showing that the upper bound of the 95% CI of C-T is below M_1 .

672

673 This is very similar to testing a superiority finding at $P \leq 0.05$. If we rule out loss of the entire
674 assumed effect of the control, we can conclude that the test drug is superior to placebo. In
675 most situations where active control studies are used, however, assuring some effect greater
676 than zero is not clinically sufficient, and the second step in selecting the NI margin is
677 choosing a specified portion of the control effect (M_1) whose loss by the test product must be
678 ruled out. This new non-inferiority margin is called M_2 , and is based upon clinical judgment.
679 The multiple steps and assumptions that are made in determining an NI margin are all
680 potential sources of uncertainty that may be introduced into the results and conclusions of an
681 NI study. This guidance attempts to identify these sources and suggest approaches to
682 accounting for these uncertainties so that we can reduce the possibility of drawing false
683 conclusions from an NI study.

684

685 Conceptually, the NI study design provides two comparisons: (1) a direct comparison of the
686 test drug with the active comparator drug, and (2) an indirect comparison of the test drug to

Contains Nonbinding Recommendations

Draft – Not for Implementation

687 placebo, based on what is known about how the effect of the active comparator compares to
688 placebo. The entire NI trial concept depends on how much is known about the size of the
689 treatment effect the active comparator will have in the NI study compared to no treatment,
690 but this effect size is not measured in the NI study and must be assumed, based on an
691 analysis of past studies of the control. The validity of the NI trial depends wholly on the
692 accuracy of the assumed effect on the control.

693

694 The assumed effect size of the active control in the NI study is based on evidence of that
695 effect derived from past trials, usually trials comparing control with placebo, but trials
696 assessing dose-response, active comparison trials, and even historically controlled trials
697 could play a role. Having assessed the effect of the active control in the past and establishing
698 HESDE (Historical Evidence of Sensitivity to Drug Effect – ICH E-10), it is then necessary
699 to decide whether that effect can be presumed to be present in the new study (the constancy
700 assumption) or must be adjusted in some way based on differences between present-day and
701 historical trials that would reduce the active control effect size. This will be discussed further
702 in section IV.B.2.d. It is also critical to ensure study quality in the NI trial, because poor
703 quality can reduce the control drug's effect size and undermine the assumption of the effect
704 size of the control agent, giving the study a "bias toward the null," which in this case
705 represents the desired outcome.

706

707 Having established a reasonable assumption for the control agent's effect in the NI study,
708 there are essentially two different approaches to analysis of the NI study, one called the *fixed*
709 *margin method* (or the two confidence interval method) and the other called the *synthesis*
710 *method*. Both approaches are discussed in later sections of section IV and use the same data
711 from the historical studies and NI study, but in different ways.

712

713 Briefly, in the fixed margin method, the margin M_1 is based upon estimates of the effect of
714 the active comparator in previously conducted studies, making any needed adjustments for
715 changes in trial circumstances. The NI margin is then pre-specified and it is usually chosen
716 as a margin smaller than M_1 (i.e., M_2), because it is usually felt that for an important endpoint
717 a reasonable fraction of the effect of the control should be preserved. The NI study is
718 successful if the results of the NI study rule out inferiority of the test drug to the control by
719 the NI margin or more. It is referred to as a fixed margin analysis because the past studies
720 comparing the drug with placebo are used to derive a single fixed value for M_1 , even though
721 this value is based on results of placebo-controlled trials (one or multiple trials versus
722 placebo) that have a point estimate and confidence interval for the comparison with placebo.
723 The value typically chosen is the lower bound of the 95% CI (although this is potentially
724 flexible) of a placebo-controlled trial or meta-analysis of trials. This value becomes the
725 margin M_1 , after any adjustments needed for concerns about constancy. The fixed margin
726 M_1 , or M_2 if that is chosen as the NI margin, is then used as the value to be excluded for C-T
727 in the NI study by ensuring that the upper bound of the 95% CI for C-T is $< M_1$ (or M_2).
728 This 95% lower bound is, in one sense, a conservative estimate of the effect size shown in
729 the historical experience. It is recognized, however, that although we use it as a "fixed"
730 value, it is in fact a random variable, which cannot invariably be assumed to represent the
731 active control effect in the NI study.

732

733 The synthesis method, derived from the same data, combines (or synthesizes) the estimate of
734 treatment effect relative to the control from the NI trial with the estimate of the control effect
735 from a meta-analysis of historical trials. This method treats both sources of data as if they
736 came from the same randomized trial, to project what the placebo effect would have been had
737 the placebo been present in the NI trial. The process makes use of the variability from both
738 the NI trial and the historical trials and yields one confidence interval for testing the NI
739 hypothesis that the treatment rules out loss of a pre-specified fixed fraction of the control
740 effect, without actually specifying that control effect or a specific fixed NI margin based on
741 the control effect.

742

743 **B. Statistical Uncertainties in the NI Study and Quantification of Treatment**
744 **Effect of Active Control**

745

746 *1. What are the Sources of Uncertainty in an NI Study?*

747

748 There are three major sources of uncertainty about the conclusions from an NI study. Two
749 of these relate to estimating the size of the effect the active control will have in the NI study
750 because that value is the basis for choosing M_1 , the non-inferiority margin whose exclusion
751 will be used to conclude that the test drug has an effect. The third is the degree of statistical
752 assurance needed in the NI study itself to determine whether the chosen NI margin has in fact
753 been ruled out.

754

755 The first source of statistical uncertainty involves the precision (or variability) of the estimate
756 of the active comparator treatment effect that is derived from an analysis of past data
757 (HESDE), whether this is based on a single randomized active comparator placebo-controlled
758 trial or from multiple trials. The uncertainty of this treatment effect estimate is quantified
759 statistically by using confidence intervals to describe the range within which the true
760 treatment effect size is likely to fall. As described in section III, assurance that the active
761 control will produce a specific effect (at least M_1) in the NI study is the single most critical
762 determination to be made in planning the NI study. Using the point estimate of the treatment
763 effect would not be an acceptable choice for the true treatment effect in the NI study because,
764 on average, half of all trials, even if the historical estimate is correct, would be expected to
765 have a smaller effect, so that one could not be reasonably sure such an effect of the control
766 was present in the NI study. It has therefore become common practice to examine the
767 confidence interval for the effect in historical experience and choose an effect that is
768 reasonably sure to be present in a new study, such as the lower bound of a 95% confidence
769 interval for the historical experience.

770

771 Particular problems arise when there is only a single historical study, as there is no
772 information about study-to-study variability (although of course, the confidence interval is
773 likely to be wider when there is only one study), when there are multiple studies but
774 substantial inconsistency in effect sizes among them, and when data from several
775 pharmacologically related drugs are used to develop the estimate for the effect of the active
776 control. When more than a single active comparator study is available, it is necessary to

Contains Nonbinding Recommendations

Draft – Not for Implementation

777 examine the results from each of the studies to determine whether the treatment effects are
778 consistent among studies or whether there are some studies where the estimate of the
779 treatment effect is zero. The need for some consistency of the active comparator effect size
780 is important and should be considered when choosing M_1 . There are also circumstances that
781 might support a less conservative choice for M_1 than the lower bound of the 95% CI for the
782 historical experience. These include factors that strongly support the expectation of a similar
783 clinical effect with the test drug, such as pharmacologic properties of the test drug that are
784 very similar to those of the active control or an effect of the test drug on a persuasive
785 biomarker.

786
787 The second source of uncertainty is not statistically based but rather arises from the concern
788 that the effect size estimated from past studies will be different from (larger than) the effect
789 of the active control in the current NI study. The need to assume that the effect will be
790 unchanged is often referred to as the “constancy assumption.” If the assumption is incorrect,
791 and the effect size in the current NI study is smaller than the estimated effect from historical
792 studies, M_1 will have been incorrectly chosen (too large) and an apparently successful study
793 showing NI could have given an erroneous result. Lack of constancy can occur for many
794 reasons, including advances in adjunctive medical care, differences in the patient
795 populations, or changes in the assessment of the endpoints under study. As noted in section
796 III, there is some experience to support the view that in outcome studies, the absolute size of
797 the treatment effect is more likely to be variable and sensitive to the background rates in the
798 control group than is the risk reduction. The risk reduction may thus be a more constant (see
799 section IV.B.2.c. on choice of metrics) measure of control drug effect than the absolute
800 effect. How to adjust the NI margin for concerns about constancy is inevitably a matter of
801 judgment.

802
803 The third source of uncertainty involves the risk of making a wrong decision from the test of
804 the non-inferiority hypothesis in the NI study (i.e., concluding that $C-T < M_1$ when it is not).
805 This uncertainty is referred to as the Type I error, or the false positive conclusion risk, and is
806 similar to the concern in a placebo-controlled trial that one might mistakenly conclude that a
807 drug is more effective than placebo. It is, in other words, present in any hypothesis-testing
808 situation. In the NI case, the statistical test is intended to ensure that the difference between
809 control and test drug ($C-T$, the degree of superiority of the control over the test drug) is
810 smaller than the NI margin, meaning that some of the effect of the control is preserved (if $C-$
811 $T < M_1$) or that a sufficient amount is preserved (if $C-T < M_2$). Typically, the one-sided
812 Type 1 error is set at 0.025, by asking that the upper bound of the 95% CI for $C-T$ be less
813 than the NI margin; this is roughly similar to the usual statistical test for a placebo-controlled
814 trial. If only one NI study is going to be conducted, the probability of a Type 1 error can be
815 made smaller by requiring that the upper bound of a CI greater than 95% be calculated and
816 be less than the margin. This is similar to what is commonly done for a single placebo-
817 controlled trial (e.g., testing at an alpha of 0.001 instead of 0.05). As noted earlier, however,
818 there may be prior information that eases this concern, and a single study at the usual Type 1
819 error boundary (0.025) may be considered sufficient if, for example, the drug and active
820 control are pharmacologically similar.

821

Contains Nonbinding Recommendations

Draft – Not for Implementation

822 This guidance will discuss the impact of the first two sources of uncertainty on the
823 quantitative approaches to estimating the control treatment effect under different assumptions
824 for these uncertainties, as well as the choice of margin to use in hypothesis testing.

825

826 2. *Quantification of the Treatment Effect of the Active Comparator*

827

828 Past controlled studies of the active control provide the empirical data for estimating the size
829 of the treatment effect of the active comparator drug. The magnitude of that treatment effect,
830 which will be the initial basis for determining the control drug effect that can be assumed to
831 be present in the NI study, is critical to determining whether conducting an NI study is
832 feasible. If the active comparator has a small treatment effect, or an effect only marginally
833 distinguished from placebo, or an inconsistent effect, an active controlled study designed to
834 show non-inferiority is likely to require a very large sample size or not be practical at all.

835

836 The magnitude of the treatment effect of the active comparator may be determined in several
837 ways, depending upon the amount of data and the number of separate studies of similar
838 design available to support this determination. The availability of many independent studies
839 is generally more informative for this determination, because the estimate of the active
840 comparator treatment effect size can be more precise and less subject to uncertainty, and
841 because it becomes possible to judge the constancy of the effect for at least the period of the
842 studies.

843

844 a. *Determining HESDE from a single study*

845

846 The most common situation in which an NI design is used involves outcome studies where
847 the active control drug has been approved for use to reduce the risk of major events (death,
848 stroke, or heart attack). It is not unusual for such approval to have been based on a single
849 study in a specific setting, although there may be other pertinent data in related conditions or
850 in different populations, or with pharmacologically similar drugs. Generally, basing an NI
851 margin on a single randomized placebo-controlled superiority study would need to take into
852 account the variability of the data in that study. The estimate of the treatment effect is
853 usually represented by some metric such as the difference between the event rate in the active
854 treatment group and the placebo control group, which can be an absolute difference in event
855 rates or a risk ratio. The treatment effect has an uncertainty that is usually measured by the
856 confidence interval, a representation of where the result is likely to be 95% of the time (for a
857 95% CI) in a future study. As a crude gauge, the lower bound of the 95% CI is
858 approximately the effect size demonstrated at a p-value of 0.025 one-sided. It is common to
859 use this value as the effect size we can be reasonably sure the active control had in the
860 historical study and is very likely to have in a future NI study. It is, on average, a low
861 estimate of the effect of the drug, and is “conservative” in that sense, but it is an effect size
862 that has a high probability of being achieved by the active control in the NI study. In
863 contrast, the point estimate of the effect seen in the historical study represents an effect size
864 that may be closer to the true effect of the active control but is one that may not be obtained
865 in a substantial fraction of any new studies. It is critical to choose the estimate of effect size
866 conservatively (i.e., one that previous studies show is very likely to be attained in the NI

Contains Nonbinding Recommendations

Draft – Not for Implementation

867 study) because the entire logic of the NI study rests on assurance that the active control in the
868 NI study has an effect size at least equal to M_1 , the largest possible NI margin.

869

870 Generally, therefore, for the fixed margin approach to setting the NI margin, the lower bound
871 of the confidence interval of the effect size of the active comparator in its historical placebo-
872 controlled experience is used to determine M_1 in order to be reasonably sure that the active
873 control will have at least the effect defined as the M_1 in the NI study. The situation improves
874 if the p-value of the estimated treatment effect is much smaller than 0.05, say in the range of
875 0.01 or 0.001 or even smaller, because in that case the lower bound of the 95% CI will
876 generally be well above zero (in absolute value) or 1.0 (for hazard ratio and other risk
877 estimates). In this case, we are more certain that the treatment effect is real and that the
878 effect of the control in the NI study will be of reasonable size.

879

880 When there is only a single trial, there is no objective assessment of study-to-study
881 variability, and there is inevitably concern about the level of assurance we can have that the
882 control will have an effect of a particular size in the NI study. A potential cautious approach
883 to account for this possible variability is to use the lower bound of a wider CI, such as the
884 99% CI. This is possible where the effect is very large, but will often yield an M_1 that
885 necessitates a very large NI trial. It may be reassuring in such cases if closely related drugs,
886 or the control drug in closely related diseases, have similar effects. A high level of internal
887 consistency in subpopulations (e.g., if the effect of the control drug is similar in subgroups
888 based on gender or age), could also provide some reassurance as to the reproducibility of the
889 result. Such findings might support use of the 95% CI lower bound even if there is only a
890 single study of the active control drug in the population to be studied in the NI trial.

891

892 b. Determining HESDE from multiple trials

893

894 Identical clinical trials in identical populations can produce different estimates of treatment
895 effect by chance alone. The extent to which two or more studies produce estimates of
896 treatment effect that are close is a function of the sample size of each study, the similarity of
897 the study populations, the conduct of the studies (e.g., dropout rates), and other factors that
898 are probably not measurable. Therefore, another source of uncertainty to be considered when
899 choosing a margin for the current NI study is the study-to-study variability in the estimate of
900 treatment effect.

901

902 When there are multiple studies of the active comparator treatment relative to a placebo or no
903 treatment, the opportunity exists to obtain an overall estimate of the active control treatment
904 effect as well as a measure of the study-to-study variability of that treatment effect. When
905 multiple studies of the active control are available, meta-analytic strategies may be used to
906 obtain a more precise estimate of the active control effects. But study-to-study variability in
907 the active comparator treatment effect is a critical consideration as well, because one of the
908 basic assumptions in NI studies is the consistency of the effect size between the historical
909 studies and the current NI study.

910

Contains Nonbinding Recommendations

Draft – Not for Implementation

911 Several special cases illustrate the use of multiple studies and problems that can arise. In
912 some of these, when the study-to-study variability is great, the need to provide assurance that
913 the control will have a definable effect size in the NI study (M_1) makes it necessary to adopt
914 a conservative estimate of the effect size.

915

916 1. The ideal case is one where there are many studies, each of sufficient size to
917 demonstrate the effect of the active control, or where there are several large outcome
918 studies, each of which has demonstrated an effect of the control, and where the effect
919 sizes derived from these studies are reasonably consistent, so that a pooled estimate,
920 obtained by a meta-analytic approach, provides a very stable and precise estimate of
921 the control effect size (narrow 95% confidence bounds) and allows a choice of M_1 that
922 is large enough to allow a reasonable choice for an M_2 margin and for the design of an
923 NI study of reasonable size.

924

925 2. If there are many small studies, where some of them have not demonstrated an effect
926 of the active control, a pooled estimate of the active control effect size and its
927 confidence interval using a random effects model can still be useful, provided there is
928 no evidence of statistical heterogeneity among the study effect sizes.

929

930 3. If there are several large outcome studies, some variation of effect sizes is expected,
931 but it would be inappropriate to have the point estimate for one of these fall below the
932 95% CI lower bound of the pooled study data, suggesting that an explanation of these
933 differences is needed and, in the absence of such an explanation, that it is not possible
934 to determine an NI margin. In this case, a clear failure of one study to show any effect,
935 again, without good explanation, such as wrong choice of endpoint or study population
936 or inadequate sample size, would also argue against the use of an NI design.

937

938 4. There are sometimes several large trials of different drugs in a pharmacologic class.
939 Pooling them may allow calculation of a 95% CI lower bound with a narrower CI that
940 yields a higher estimate of the active control drug effect than would any single study.
941 The presumption that the pharmacologically similar drugs would have similar effects
942 may be reasonable, but care should be exercised in extending this assumption too far.

943

944 If the effect size of these different drugs varies considerably in the trials, it may be
945 reasonable to use the pooled data to estimate effect size, but it appears desirable to use
946 the drug with the largest effect (point estimate) as the active control in the NI study,
947 even if the pooled data (95% CI lower bound) are used to estimate the active control
948 effect size.

949

950 When an analysis is based on multiple studies, it is important to consider all studies and all
951 patients. Dropping a study that does not show an effect, unless there is a very good reason,
952 can overestimate the control drug effect and give a falsely high M_1 . As noted above, the
953 existence of properly designed and sized studies that show no treatment effect of the active
954 comparator may preclude conducting NI studies with that active comparator unless there are
955 valid reasons to explain these results.

956

957 Examples 1, 3, and 4 in the Appendix illustrate in more detail how multiple historical
958 placebo-controlled trials of the active comparator studies are evaluated.

959

960

c. Metrics of treatment effect

961

962 There are several different metrics that can be used to assess the treatment effect estimated in
963 an NI study. These include the following:

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981 The metric used in calculating HESDE need not be the one used in the original study. If
982 placebo response rates differ markedly among several studies in a meta-analysis, it is
983 generally more sensible to analyze relative risk than absolute risk. It seems far more likely
984 that in the NI study it will be the risk reduction, not the absolute effect, that will be constant.

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

Note that we are using the convention that for the ratio of risks (bad outcomes such as failure rates or deaths) in the historical trials, risks are shown as control drug/placebo (i.e., the drug is the numerator), so that the RR (or HR) will be less than 1. In an NI study, the control drug becomes the denominator and the test drug is the numerator, with a risk increase to be ruled out. For example, if the control gives a 25% risk reduction relative to placebo, what must be ruled out to show that the NI margin is excluded is an increased risk of 33%, or an RR of

Contains Nonbinding Recommendations

Draft – Not for Implementation

1001 1.33, calculated by dividing the active drug effect versus placebo into 1 ($1 \div 0.75 = 1.33$). How
1002 to calculate M_2 is not entirely straightforward. If we take half of the control effect versus
1003 placebo, for an HR of 0.875, then convert that to the risk increase to be ruled out, we get
1004 $1 \div 0.875$ or 1.14. If, on the other hand, we take half of the 33% increase calculated earlier,
1005 we get 1.165.

1006
1007 Whether to calculate M_2 before or after changing numerator and denominator is not settled.
1008 A way to calculate the margin without this asymmetry is to convert the HR to the natural
1009 logarithm scale. When the natural logarithm transformation of the risk ratio is used, that is,
1010 $\log(A/B)$ and $\log(B/A)$, the two logs have the same magnitude except that the signs are
1011 opposite. In the previous example, for 50% retention of the 25% risk reduction in the NI
1012 study, the non-inferiority margin for $\log(T/C)$ is the mid-point between $\log(4/3)$ and zero.
1013 By converting log risk ratio back to risk ratio, the non-inferiority margin for T/C is the square
1014 root of 4/3, giving a value of 1.155. The margin calculated that way then falls between the
1015 1.14 and 1.165 calculated previously.

1016
1017 The difference between expressing the treatment effect as the absolute difference between
1018 success rates in treatment groups and as the relative risk or risk ratio for success on the test
1019 treatment relative to the active comparator is illustrated in the following two examples.

1020
1021 For the first example, consider a disease where the cure rate is at least 40% in patients
1022 receiving the selected active control and 30% for those on placebo, a 10% difference in cure
1023 rates. If the purpose of an NI study is to demonstrate that the test product is effective (i.e.,
1024 superior to a placebo), then the difference between the test product and active control in the
1025 NI study must be less than 10%. The margin M_1 would then be 10%. If the additional
1026 clinical objective is to establish that the test product preserves at least half of the active
1027 control's effect, then the cure rate of the test product must be shown to be less than 5% worse
1028 than the control, the M_2 margin.

1029
1030 This approach depends on the control drug's having an effect of at least 10% greater than a
1031 placebo (had there been one) in the NI study. If the population in the NI study did not have
1032 such a benefit (e.g., if the patients all had viral illnesses such that the benefit was less than
1033 10%), then even if the 5% difference were ruled out, that would not demonstrate the desired
1034 effectiveness (although it would seem to). Note that in this case, if the true effect of the
1035 control in the study were 8%, then ruling out a 5% difference would in fact show some effect
1036 of the test drug, just not the desired 50% of control effect.

1037
1038 The second example illustrates a non-inferiority margin selected for the risk ratio
1039 (test/control) metric. Let C and P represent the true rates of an undesirable outcome for the
1040 control and a placebo, respectively. The control's effect compared to placebo is expressed by
1041 the risk ratio, C/P. A risk ratio of 1 represents no effect; a ratio of less than 1 shows an
1042 effect, a reduction in rate of undesirable outcomes.

1043
1044 Metrics like the risk ratio may be less affected by variability in the event rates in a placebo
1045 group that would occur in a future study. For example, a risk ratio for the event of interest of

Contains Nonbinding Recommendations

Draft – Not for Implementation

1046 3/4 =0.75 can be derived from very different absolute success results from different studies,
1047 as shown in the table below. While the risk ratio is similar in all four hypothetical studies,
1048 the absolute difference in success rates ranges from 5% to 20%. Suppose that the NI margin
1049 were based on historical studies showing control drug effects like those in the fourth study.
1050 The NI margin would then be chosen as 20%. Now suppose that under more modern
1051 circumstances the NI study had a control rate more like Study 1 and an effect size vs. placebo
1052 of far less than 20%. An NI margin (M_1) of 20% would then be far greater than the drug
1053 effect in the NI study, and ruling out a difference of 20% would not demonstrate
1054 effectiveness at all. Thus, if the NI margin were chosen as ruling out an inferiority of 33%
1055 (or a relative risk of 1.33, i.e., $1 \div 0.75$), if the control rate were 15%, the difference (M_1)
1056 between test and control would need to be less than 5% ($15\% \times 1.33 = 20\%$, or $5\% >$ the
1057 15% rate in the active control group).
1058

Study Number	Risk Ratio (C/P)	Control rate	Placebo rate
Study 1	3/4	15%	20%
Study 2	3/4	30%	40%
Study 3	3/4	45%	60%
Study 4	3/4	60%	80%

1059
1060 In this case, where absolute effect sizes vary but risk reductions are reasonably constant, the
1061 risk ratio metric provides a better adjustment to the lower event rate in the NI study.
1062

1063 These examples illustrate the importance of understanding how a particular metric will
1064 perform. The choice between a relative metric (e.g., risk ratio) and an absolute metric (e.g., a
1065 difference in rates) in characterizing the effects of treatments may also be based upon clinical
1066 interpretation, medical context, and previous experience with the behavior of the rates of the
1067 outcome.
1068

1069 d. The Concept of “Discounting” the Treatment Effect Size to Account for 1070 Various Sources of Uncertainty 1071

1072 One of the strategies employed in choosing the margin M_1 for the NI study design is that of
1073 “discounting” or reducing the magnitude of the margin size that is used in the NI study from
1074 what is calculated from the analysis of HESDE. Such discounting is done to account for the
1075 uncertainties in the assumptions that need to be made in estimating, based on past
1076 performance, the effect of the active control in the NI study. This concept of discounting
1077 focuses on M_1 determination and is distinct from a clinical judgment that the effect that can
1078 be lost on clinical grounds should be some fraction of M_1 (i.e., M_2). As discussed above,
1079 there are uncertainties associated with translating the historical effect of the active control
1080 (HESDE) to the new situation of the active control NI trial, and it is tempting to deal with
1081 that uncertainty in the constancy assumption by discounting the effect (“take half”). To the
1082 extent possible, concerns about the active control effect should be as specific as possible,
1083 should use available data (e.g., magnitude of possible differences in effect in different patient
1084 population, consistency of past studies, and consistency within studies across population
1085 subsets should be examined), and should take into account factors that reduce the need for a

1086 conservative estimate, such as the pharmacologic similarity of the test and control drugs and
1087 pharmacodynamic effects of the new drug, rather than reflecting “automatic” discounting.
1088 Having considered these matters, if significant uncertainties remain, an approach that further
1089 discounts or reduces, say by 25%, the magnitude of the active control effect based on
1090 HESDE can be considered.

1091

1092 A closely related issue is adjustment of M_1 to reflect a finding that the population in the NI
1093 study was different from the historical study in such a way that what the historical experience
1094 shows would lead to a smaller effect size (e.g., a finding of a smaller effect in women would
1095 need to be considered in assessing the validity of M_1 if the NI study had substantially more
1096 women than the historical studies). In general, the assessment of the historical data should
1097 identify such differences so that plans for the NI study take this into account or so that the
1098 value of M_1 can be revisited in light of the study population included in the NI study.

1099

1100 **C. Statistical Methods for NI Analysis**

1101

1102 Several approaches are used to demonstrate statistically that the NI objective is met. Each
1103 statistical approach to demonstrating NI depends upon a number of factors including:

1104

- 1105 • What assumptions are made and how verifiable or empirically demonstrable these
1106 assumptions are
- 1107 • The degree to which judgment, both statistical and clinical, is exercised in accounting
1108 for the various uncertainties in the data from the current NI study and also from the
1109 clinical trials of the active control that are the basis for estimating its effect
- 1110 • The clinical judgment of how much of the treatment effect of the active comparator
1111 can be lost (M_2 selection)

1112

1113 As noted earlier, the two main approaches to demonstrating non-inferiority are the fixed
1114 margin method and the synthesis method.

1115

1116 Each of these statistical approaches uses the same data from the previously conducted
1117 controlled trials of the active control and the same data from the current NI study, but the
1118 approaches are different in several ways. The first is with regard to their emphasis on the
1119 specific determination for M_1 before determining M_2 . There is also a difference between
1120 them in how the data from the historical studies and the NI study are used or combined.

1121 What follows is a guide to the differences between the two approaches. Examples 1(A) and
1122 1(B) in the Appendix provide more detailed illustrations of how each of these approaches is
1123 used and interpreted. In general, the fixed margin approach is more conservative and treats
1124 the variance of the NI study and historical evidence distinctly. That is, a very large historical
1125 database will give a narrower CI and larger 95% lower bound for M_1 , but it will not directly
1126 figure into the test drug versus placebo calculation, as is done in the synthesis method.

1127 Concern about using the synthesis approach reflects our view that the method incorporates
1128 too much certainty about the past results into the NI comparison. We believe the fixed
1129 margin approach is preferable for ensuring that the test drug has an effect greater than

Contains Nonbinding Recommendations

Draft – Not for Implementation

1130 placebo (i.e., the NI margin M_1 is ruled out). However, the synthesis approach, appropriately
1131 conducted, can be considered in ruling out the clinical margin M_2 .

1132

1133 1. *The Fixed Margin Approach for Analysis of the NI Study*

1134

1135 Sections IV.B.2.a and B.2.b contain discussions of the basic statistical approach to estimating
1136 the active comparator treatment effect size from past controlled trials. The goal of these
1137 analyses is to define the margin M_1 , a fixed value, based on the past effect of the active
1138 control, which is intended to be no larger than the effect the active control is expected to have
1139 in the NI study. Whether M_1 is based on a single study or multiple studies, the observed (if
1140 there were multiple studies) or anticipated (if there is only one study) statistical variation of
1141 the treatment effect size should contribute to the ultimate choice of M_1 , as should any
1142 concerns about constancy. The selection of M_2 is then based on clinical judgment regarding
1143 how much of the M_1 active comparator treatment effect can be lost. The exercise of clinical
1144 judgment for the determination of M_2 should be applied after the determination of M_1 has
1145 been made based on the historical data and subsequent analysis.

1146

1147 All relevant studies of the active comparator and all randomized patients within these studies
1148 should generally be used in determining the margin M_1 because that provides a more reliable
1149 and, possibly, conservative estimate. The actual selection of which studies are used in a
1150 meta-analysis and how that selection is made can be complex and itself subject to judgment.
1151 See Examples 1(A), 3, and 4 that illustrate these points in the Appendix.

1152

1153 The design and analysis of the NI study, and its analysis using the fixed margin approach, is
1154 well known and described in ICH E9, section 3.3.2. This statistical approach relies upon the
1155 choice of a fixed non-inferiority margin that is pre-specified and part of the NI design. There
1156 is very little, however, in ICH E9 or ICH E10 that discusses just how to determine the
1157 margin. Although the constancy assumption and study quality issues are recognized, there is
1158 little discussion about how to adjust the margin because of such statistical or study data
1159 uncertainties. Any discounting of the historical evidence of the effect of the active control
1160 based on uncertainty of the constancy of the effect (e.g., because of changes in practice or
1161 concomitant treatment), which is an attempt to improve the estimate of the control effect in
1162 the NI study, affects the M_2 as well, as in most cases M_2 is a fraction of M_1 . M_2 might not be
1163 affected when it is very small compared to M_1 , as is the case in considering very effective
1164 drugs. It is critical to note that M_2 is a judgment that is made after M_1 is chosen, but M_2 , of
1165 course, can never be larger than M_1 . It is perhaps tempting to make up for uncertainty in M_1
1166 by demanding assurance of preservation of a larger fraction of M_1 by ruling out a smaller
1167 loss of effect (i.e., using a smaller M_2), but the temptation should be avoided. The first and
1168 most critical task in designing an NI study is obtaining the best estimate of the effect of the
1169 active control in the NI study (i.e., M_1).

1170

1171 Operationally, the fixed margin approach usually proceeds in the following manner. The
1172 active comparator effect size is calculated from past placebo-controlled studies. The lower
1173 bound of the confidence interval describing the effect of the active control in past studies, a
1174 single number, is selected as a conservative choice for the active comparator effect size.

Contains Nonbinding Recommendations

Draft – Not for Implementation

1175 While traditionally the 95% confidence interval is used, there can be flexibility in this choice,
1176 such as a 90% confidence interval or even narrower, when the circumstances are appropriate
1177 to do so (e.g., strong evidence of a class effect, strong biomarker data). It is recognized that
1178 use of a fixed margin to define the control response is conservative as it picks a “worst case”
1179 out of a confidence interval that consists of values of effect that are all larger. This choice,
1180 however, is one response to the inherent uncertainty of estimates based on past studies,
1181 including the variability of those past estimates, and the possibility that changes in medical
1182 practice, or hard to recognize differences between the past studies and the current NI study,
1183 have made the past effect an overestimate of the active control effect in the new study.

1184
1185 Although some of the uncertainty about applicability of past results to the present is reflected
1186 in a conservative choice of margin (95% of CI lower bound) used to initiate consideration of
1187 M_1 , there may be further concerns about past variability and constancy that lead to a
1188 determination to discount this lower bound further in choosing M_1 to account for any sources
1189 of uncertainty and dissimilarities between the historical data and the NI study to be
1190 conducted, as discussed in the earlier sections. Following this, a clinical judgment is made as
1191 to how much of this effect should be preserved. This clinical judgment could choose M_2 to
1192 be the same as M_1 , but as noted, where the treatment effect is important (e.g., an effect on
1193 mortality) it is usual to ask that a reasonable fraction of the control effect be preserved, by
1194 making M_2 , the loss of effect to be ruled out, smaller than M_1 . Choosing M_2 as 50% of M_1
1195 has become usual practice for cardiovascular (CV) outcome studies, whereas in antibiotic
1196 trials, where effect sizes are relatively large, a 10-15% NI margin for M_2 is common. Note
1197 that the M_2 of 50% of M_1 is on a relative scale, whereas the 10-15% is on the absolute scale
1198 for antibiotic drugs. The analysis of the NI study involves only the data from the NI study,
1199 and the test of the hypothesis that inferiority greater than the M_2 margin has been excluded is
1200 statistically similar to showing that the 95% CI in a superiority study excludes a difference of
1201 zero.

1202
1203 Thus, there are two confidence intervals involved in the fixed margin approach, one from the
1204 historical data, where one uses the lower bound to choose M_1 , and one from the NI study (to
1205 rule out $C-T > M_2$); in this example both intervals are 95% confidence intervals. That is why
1206 this fixed margin approach is sometimes called the 95%-95% method. It should be
1207 appreciated that the analysis of the NI study (ruling out a difference $> M_2$ by examining the
1208 lower bound of the CI for C-T) is the analysis that is based on the randomized comparison in
1209 the NI study, in contrast to the determination of M_1 , which is not based on a concurrent
1210 randomization.

1211
1212 Separating the process of estimating the treatment effect of the active comparator based upon
1213 the historical data (i.e., choice of M_1) from the analysis of the NI study has some advantages
1214 and disadvantages. Two important advantages are that it provides a single number that is
1215 clinically understandable for an M_1 (and derived M_2) and that it provides a basis for planning
1216 the sample size of the NI study to achieve statistical control of Type 1 error and the power
1217 needed for the NI study to meet its objective for the pre-specified NI margin. One arguable
1218 disadvantage is that the method is statistically not efficient because it uses the two confidence
1219 interval approach rather than a combined estimate of the statistical variability of the historical

Contains Nonbinding Recommendations

Draft – Not for Implementation

1220 and NI study data. Nevertheless, use of the fixed margin is readily understood, particularly
1221 by non-statisticians, and is only somewhat conservative compared to an analysis using the
1222 synthesis approach. Decisions to discount the M_1 further or, where appropriate, to use a
1223 narrower confidence interval, are easily explained, and can make the fixed margin approach
1224 more or less conservative.

1225

1226 Deciding on the NI clinical margin M_2 is also a relatively straightforward concept. It is
1227 plainly a matter of judgment about how much of the treatment effect must be shown to be
1228 preserved, a consideration that may reflect the seriousness of the outcome, the benefit of the
1229 active comparator, and the relative safety profiles of the test and comparator. It also has
1230 major practical implications. In large cardiovascular studies, it is unusual to seek retention of
1231 more than 50% of the control drug effect even if this might be clinically reasonable, because
1232 doing so will usually make the study size infeasible.

1233

1234 The fixed margin approach considers the NI margin as a single number, fixed in advance of
1235 the NI study. The hypothesis tested in the NI study determines whether the comparison of
1236 the test drug to the active control meets the specified NI criterion, assuming, of course, that
1237 the active control had at least its expected effect (equal to M_1) and that the study therefore
1238 had assay sensitivity. A successful NI conclusion, ruling out a difference $> M_1$, shows that
1239 the test drug is effective (just as a superiority study showing a significant effect at $p \leq 0.05$
1240 does) and, if a difference $> M_2$ is also ruled out, shows that the new drug preserves the
1241 desired fraction of the control drug's effect. This statistical test of hypothesis is not formally
1242 directed at determining whether the test drug would have been superior to a placebo, had a
1243 placebo group been included in the NI study, but it leads to a similar conclusion by ruling out
1244 the possibility that the test drug is inferior to the control by more than an amount equal to the
1245 whole effect of the control compared to placebo (that effect being known from past studies).

1246

1247 The possible outcomes of such trials are shown in Figures 2 and 3 in section III of this
1248 guidance.

1249

1250 *2. The Synthesis Approach for Analysis of NI*

1251

1252 An alternative statistical approach is known as the synthesis approach because it combines or
1253 synthesizes the data from the historical trials and the current NI trial, reflecting the variability
1254 in the two data sets (the current NI study and the past studies used to determine HESDE).

1255 The synthesis method is designed to directly address the question of whether the test product
1256 would have been superior to a placebo had a placebo been in the NI study, and also to
1257 address the related question of what fraction of the active comparator's effect is maintained
1258 (the loss to be ruled out) by the test product. In the synthesis approach, the NI margin is not
1259 predetermined, but the outcome of the NI study, a consideration of the effect of the test agent
1260 vs. placebo, can be judged for adequacy.

1261

1262 Although the synthesis approach combines the data from the historical trials into the
1263 comparison of the concurrent active comparator and the test drug in the NI study, a direct
1264 randomized concurrent comparison with a placebo is of course not possible, as the placebo

1265 group is not a concurrent control and there is no randomization to such a group within the NI
1266 study. The imputed comparison with a placebo group that is not in the NI study thus rests on
1267 the validity of several assumptions, just as the fixed margin approach does. The critical
1268 assumption of the constancy of the active control effect size derived from the historical
1269 controlled trials is just as important when the synthesis method is used.

1270

1271 Because of the way the variance of the historical data and the NI data are combined for the
1272 synthesis test, the synthesis test is more efficient (uses a smaller sample size or achieves
1273 greater power for the same sample size) than the fixed margin approach but requires
1274 assumptions that may not be appropriate. The statistical efficiency of the synthesis approach
1275 derives primarily from how the standard error of the comparison of test product to active
1276 comparator is dealt with. See Appendix, Example 1(B), for a comparison of the two methods
1277 and the variance calculations.

1278

1279 The synthesis approach does not specify a fixed NI margin. Rather, the method combines (or
1280 synthesizes) the estimate of treatment effect relative to the control from the NI trial with the
1281 estimate of the control effect from a meta-analysis of historical trials. The method treats both
1282 sources of data as if they came from the same randomized trial, to project where the placebo
1283 effect would have been had the placebo been present in the NI trial. The synthesis process
1284 makes use of the variability from the NI trial and the historical trials and yields one
1285 confidence interval for testing the NI hypothesis that the treatment preserves a fixed fraction
1286 of the control effect, without actually specifying that control effect or a specific fixed NI
1287 margin based on the control effect. Clinical judgment is used to pre-specify an acceptable
1288 fraction of the control therapy's effect that should be retained by the test drug, regardless of
1289 the magnitude of the control effect.

1290

1291 A disadvantage of the synthesis approach, however, is that it does not allow for a pre-
1292 specification of the actual size or magnitude of the NI margin M_1 , so the clinical judgment to
1293 determine the choice of M_2 is difficult and is generally not made until results are seen.
1294 Moreover, it may be unrealistic to assign the same weight to the variance of the historical
1295 outcome data and to that of the concurrent randomized NI treatment. As also noted, the
1296 efficiency of the fixed margin approach can sometimes be enhanced either formally, by
1297 including more trials (e.g., of related drugs) in the historical meta-analysis, and thereby
1298 increasing the margin M_1 , or, as a matter of judgment, by considering pharmacologic
1299 similarities between the control and test drugs, effects on pertinent biomarkers (e.g., tumor
1300 response rate), all of which could lead to choice of a fixed margin based on a less extreme
1301 boundary of the confidence interval (e.g., 80% instead of 95%).

1302

1303 **D. Considerations for Selecting M_2 , the Clinical Margin, and the Role of** 1304 **Subjective Judgment**

1305

1306 M_2 is the margin that is the pre-specified NI margin that should be met in an NI study. The
1307 determination of M_2 is based on clinical judgment and is usually calculated by taking a
1308 percentage or fraction of M_1 . The clinical judgment in determining M_2 may take into account
1309 the actual disease incidence or prevalence and its impact on the practicality of sample sizes

1310 that would have to be accrued for a study. There can be flexibility in the M_2 margin, for
1311 example, when:

1312

- 1313 (1) The difference between the active comparator response rate and the spontaneous
1314 response rate is large;
- 1315 (2) The primary endpoint does not involve an irreversible outcome such as death (in
1316 general, the M_2 margin will be more stringent when treatment failure results in an
1317 irreversible outcome);
- 1318 (3) The test product is associated with fewer serious adverse effects than other therapies
1319 already available;
- 1320 (4) The test product is in a new pharmacologic category and has been shown to be
1321 tolerated by patients who do not tolerate therapies that are already available.

1322

1323 There is also a difference in implication when the study NI conclusion is “not quite”
1324 significant (M_1 is not excluded) for M_1 and when this is the case for M_2 . Failure to exclude
1325 inferiority equal to M_1 means there is no assurance of any effect. Just as, for a placebo-
1326 controlled trial, it would be most unusual to accept as positive a study with $p > 0.05$, it would
1327 be most unusual to accept an NI study where the upper bound of 95% CI was $> M_1$. On the
1328 other hand, failing to exclude M_2 by a small amount means that instead of ruling out a loss of
1329 50% of M_1 , you have ruled out, say, a 48% loss, not necessarily a definitive failure. As noted
1330 above, we would also consider the less conservative synthesis approach in assessing M_2 .

1331

1332 **E. Estimating the Sample Size for an NI Study**

1333

1334 It is important to plan the sample size for an NI clinical trial so that the trial will have the
1335 statistical power to conclude that the NI margin is ruled out if the test drug is truly non-
1336 inferior. This is not always an easy task. At the protocol planning stage, using the fixed
1337 margin approach, the magnitude of the NI margin will be specified; the sample size must be
1338 based on the need to rule out inferiority greater than M_2 . This should usually be based on an
1339 NI using a fixed margin approach. The margin to be ruled out is the most critical component
1340 of the sample size planning, but the variance of the estimate of the treatment effects will not
1341 be known and it is also critical. A further problem is posed by the possibility that event rates
1342 will be lower in the new study. In this case, if the NI margin is expressed as, for example,
1343 ruling out (at the upper bound of the 95% CI for C-T) an increase in risk of 25%, this will be
1344 far easier when the event rate on active control is 8% than when it is 4%, even if the active
1345 control is superior to placebo by the same absolute 20% difference. This problem is not
1346 different from specifying sample size in a superiority trial. It too depends on the event rate,
1347 and it is common to examine blinded data during the trial to see if the event rate is
1348 unexpectedly low. A similar approach could be applied in an NI trial with upward
1349 adjustment of the sample size if the event rate is unexpectedly low. There is one further
1350 consideration. If, in reality, the test drug is somewhat more effective than the control, it will
1351 be easier to rule out any given NI margin and a smaller sample size could be used. A
1352 somewhat less effective test drug will, of course, require a larger sample size.

1353

1354 **F. Potential Biases in an NI Study**

1355

1356 Traditionally, analysis of the results of randomized clinical superiority trials follows the
1357 intent-to-treat principle, namely, that all randomized patients are analyzed according to the
1358 treatment to which they were randomized. This analysis is intended to avoid various biases
1359 associated with patients switching treatment, selection bias, and dropout/withdrawal patterns
1360 that may confound the observed treatment effect. This is recognized as a potentially
1361 conservative analysis. Including patient outcomes that occur after a patient has stopped the
1362 treatment, for example, or show poor compliance with treatment, would be expected to bias
1363 the analysis toward the null (no treatment difference). Intent-to-treat (ITT) analyses in
1364 superiority trials are nonetheless preferred because they protect against the kinds of bias that
1365 might be associated with early departure from the study. In non-inferiority trials, many kinds
1366 of problems fatal to a superiority trial, such as non-adherence, misclassification of the
1367 primary endpoint, or measurement problems more generally (i.e., “noise”), or many dropouts
1368 who must be assessed as part of the treated group, can bias toward no treatment difference
1369 (success) and undermine the validity of the trial, creating apparent non-inferiority where it
1370 did not really exist. Although an “as-treated” analysis is therefore often suggested as the
1371 primary analysis for NI studies, there are also significant concerns with the possibility of
1372 informative censoring in an as-treated analysis. It is therefore important to conduct both ITT
1373 and as-treated analyses in NI studies. Differences in results using the two analyses will need
1374 close examination. The best advice for conducting an NI study is to be aware at the planning
1375 stage of these potential issues and to monitor the trial in a manner that minimizes these
1376 problems, as they can seriously affect the validity of an NI study.

1377

1378 Other sources of bias that could occur in any study are also of concern in the NI study and
1379 are of particular concern in an open label study. For such open label NI studies, how best to
1380 ensure unbiased assessment of endpoints, unbiased decisions about inclusion of patients in
1381 the analysis, and a wide variety of other potential biases, need particular attention.

1382

1383 **G. Role of Adaptive Designs in NI Studies — Sample Size Re-estimation to**
1384 **Increase the Size of an NI Trial**

1385

1386 Because it may be difficult to adequately plan the sample size for any study, including an NI
1387 study, especially when assumptions like the event rate may change from the planning phase
1388 to the study conduct, adaptive study designs that can allow for the prospective re-estimation
1389 of a larger sample size can be considered. The most critical single consideration in such
1390 designs is precise knowledge about whether there is unblinding as to treatment. Sample size
1391 re-estimation, if based on a blinded analysis of the overall variance estimate or the overall
1392 event rate, without knowledge of or a comparison of the unblinded treatment group response
1393 rates or the differences between treatment groups, is not only acceptable but generally
1394 advisable. It is critical to provide reassurance and procedures that ensure maintenance of
1395 blinding.

1396

1397 If an adaptive design that allows unblinding is contemplated, then the design features and
1398 procedures for protection of the integrity of the trial need to be clearly stated in the protocol

1399 for the trial. Some adaptive designs may include an independent Data Monitoring
1400 Committee (DMC) to monitor the planned adaptation. The DMC charter should address
1401 procedures for the sharing and blinding of data, and the procedures used to maintain a
1402 firewall between those who do, and those who do not view unblinded data. Some of these
1403 issues will be addressed in a companion guidance on Adaptive Study Designs.

1404

1405 **H. Testing NI and Superiority in an NI Study**

1406

1407 In general, when there is only one endpoint and one dose of the test treatment, a planned NI
1408 study can be tested for superiority without a need for Type 1 error alpha correction. That is,
1409 the same 95% or higher confidence interval employed for testing non-inferiority with the pre-
1410 specified fixed margin can be used to test superiority. One can also think of this as a two-
1411 stage analysis in which the showing of NI using a 95% confidence interval (invariably
1412 successful if the test drug is actually superior), is then followed sequentially by superiority
1413 testing. This sequential testing has the Type I error rates for both non-inferiority and
1414 superiority controlled at a level of no more than 5%. A non-inferiority showing after a failed
1415 superiority study, in contrast, gives a generally uncertain result, and such a study would
1416 generally be considered a failed study. Thus, successful showing of non-inferiority allows
1417 superiority testing but a failed showing of superiority would yield credible evidence of non-
1418 inferiority only if the study were designed as a non-inferiority study (e.g., the NI margin must
1419 be pre-specified, and assay sensitivity and HESDE must be established).

1420

1421 When there are multiple endpoints or multiple doses of the test treatment evaluated in an NI
1422 study, the valid statistical decision tree can be very complex. Using the same 95%
1423 confidence interval to test non-inferiority and superiority at each endpoint level or at each
1424 dose may inflate the overall Type I error rate associated with drawing one or more false
1425 conclusions from such multiple comparisons, regardless of whether they are non-inferiority
1426 or superiority testing. Thus, for any statistical decision tree composed of tests of superiority
1427 and non-inferiority in multiple comparison settings, it is imperative to evaluate the overall
1428 Type I error rate for all the comparisons involved in the testing and make appropriate
1429 statistical adjustments.

1430

1431 Some of the problems in interpreting the results of non-inferiority analyses are more subtle
1432 than those with superiority testing. In particular, as noted previously, design or conduct
1433 problems such as medication non-compliance or misclassification/measurement error, errors
1434 that would be fatal to success in a superiority study, can lead to apparently favorable (results)
1435 in a non-inferiority study.

1436

1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481

V. COMMONLY ASKED QUESTIONS AND GENERAL GUIDANCE

1. Can a margin be defined when there are no placebo-controlled trials for the active control for the disease being assessed?

If the active control has shown superiority to other active treatments in the past, the difference demonstrated represents a conservative estimate of HESDE, one that can certainly serve as a basis for choosing M_1 . It may also be possible that trials of the active control in related diseases are relevant. The more difficult question is whether historical experience from nonconcurrently controlled trials can be used to define the NI margin. The answer is that it can, but the circumstances are similar to those in which a historically controlled trial can be persuasive (see ICH E-10). First, there should be a good estimate of the historical spontaneous cure rate or outcome without treatment. Examination of medical literature and other sources of information may provide data upon which to base these estimates (e.g., historical information on natural history or the results of ineffective therapy). Second, the cure rate of the active control should be estimated from historical experience, preferably from multiple experiences in various settings, and should be substantially different from the untreated rate. For example, if the spontaneous cure rate of a disease is 10-20% and the cure rate with an active control is 70-80%, these are substantially different and an acceptable margin, generally chosen conservatively, can probably be identified for M_1 . The clinically acceptable loss of this effect can then be determined for M_2 . Estimates of the cure rate of the active control should be based upon data from clinical trials, even if these are not controlled, and it is critical to be sure the trial patients and untreated patients are similarly defined and selected. Example 2 in the Appendix illustrates a case of this kind, in which it was concluded that a margin could be defined despite the absence of placebo-controlled trials of the active control. It becomes more difficult to identify a margin when the difference between the spontaneous cure rate and active drug cure rate is smaller. For example, if the historical spontaneous cure rate is 40% and the active control rate is 55%, it would not be credible to identify the NI margin in this case as 15%, as such a small difference could easily be the result of different disease definition or ancillary therapy. When the historical cure rates for the active control and the cure rate in patients who receive no treatment are not known at all from actual studies (i.e., are just based on clinical impressions), it will be difficult or impossible to define an NI margin.

2. Can the margin M_2 be flexible?

As indicated in sections III and IV, there is a critical difference between demonstrating in the NI study that the margins M_1 and M_2 have been met. M_1 is used to determine whether the NI study shows that the test drug has any effect at all. Accepting a result in which the 95% CI did not rule out loss of M_1 would be similar to accepting, as showing effectiveness, a superiority study whose estimated treatment effect was not significant at $p \leq 0.05$. M_2 , in contrast, represents a clinical judgment about what level of loss of the active control effect is acceptable. A typical value for M_2 is often 50% of M_1 , at least

1482 partly because the sample sizes needed to rule out a smaller loss become impractically
1483 large. In this case, there is a better argument for some degree of flexibility if the study
1484 did not quite rule out the M_2 margin; there might be reason to consider, for example,
1485 assurance of 48% retention (but not the expected 50%) for M_2 as acceptable. We have
1486 also concluded that the fixed margin method, more conservative but with fewer
1487 assumptions, should generally be used in ensuring that loss of M_1 is ruled out but that the
1488 synthesis method can be used to assess M_2 . Of course, allowing too much inferiority of
1489 the test drug to the standard, especially for endpoints of mortality and serious morbidity,
1490 would clearly not be acceptable.

1491

1492 **3. Can prior information or other data (e.g., studies of related drugs, pharmacologic**
1493 **effects) be considered statistically in choosing the NI margins or in deciding whether**
1494 **the NI study has demonstrated its objective?**

1495

1496 Prior information could be characterized in a statistical model or in a Bayesian
1497 framework by taking into account such factors as evidence of effects in multiple related
1498 indications or on many endpoints. Such information might be used in determining M_1 in
1499 a more flexible (less conservative) manner. For example, if multiple studies provide very
1500 homogeneous results for one or more important endpoints it may be possible to use the
1501 90% lower bound rather than the 95% lower bound of the CI to determine the active
1502 control effect size. Similarly, if there were additional supporting evidence for the clinical
1503 effect of the test drug, such as prior information on the efficacy of the test drug in related
1504 diseases or in a compelling animal model, or an effect on an important biomarker (e.g.,
1505 tumor response rate), or evidence that pharmacologically related drugs were clearly
1506 effective in the condition being studied, such prior information would increase the
1507 evidence for the plausibility of the intended NI effect of the test drug, which might allow
1508 use of a less conservative estimate of effect than the 95% lower bound of the confidence
1509 interval for C-T in the NI study. Finally, a statistical model such as a regression
1510 adjustment may be applied to the NI study analysis if the covariates for patients in the
1511 historical clinical studies are distributed differently from those of patients in the current
1512 NI study. This adjustment may, in some situations, reduce the variance of the NI test and
1513 increase the ability of the comparison to meet the NI margin. In other situations, where
1514 there is more heterogeneity of the covariates, the variance may be increased, adversely
1515 impacting the comparison.

1516

1517 **4. Can a drug product be used as the active comparator in a study designed to show**
1518 **non-inferiority if its labeling does not have the indication for the disease being**
1519 **studied, and could published reports in the literature be used to support a treatment**
1520 **effect of the active control?**

1521

1522 The active control does not have to be labeled for the indication being studied in the NI
1523 study, as long as there are adequate data to support the chosen NI margin. FDA does, in
1524 some cases, rely on published literature and has done so in carrying out the meta-analyses
1525 of the active control used to define NI margins. An FDA guidance for industry on
1526 *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*

1527 describes the approach to considering the use of literature in providing evidence of
1528 effectiveness, and similar considerations would apply here. Among these considerations
1529 are the quality of the publications (the level of detail provided), the difficulty of assessing
1530 the endpoints used, changes in practice between the present and the time of the studies,
1531 whether FDA has reviewed some or all of the studies, and whether FDA and the sponsor
1532 have access to the original data. As noted above, the endpoint for the NI study could be
1533 different (e.g., death, heart attack, and stroke) from the primary endpoint (cardiovascular
1534 death) in the studies if the alternative endpoint is well assessed (see also question 6).
1535

1536 **5. If the active control drug is approved for the indication that is being studied, does**
1537 **the margin need to be justified, or if the active control drug has been used as an**
1538 **active comparator in the past in another study of design similar to the current study**
1539 **and a margin has been justified previously, can one simply refer to the previous**
1540 **margin used?**

1541
1542 When an active control drug is approved, the effect size for the indication is not usually
1543 identified in a pooled analysis, nor is the variability of that effect size in the various trials
1544 calculated. It would therefore be difficult to base the NI margin on the label of the active
1545 control drug. On the other hand, FDA’s reliance on the studies for approval would
1546 support the view that the quality of the studies was acceptable and that the studies could
1547 contribute to a determination of the NI margin. In general, approval of a drug is based on
1548 showing superiority to placebo, usually in at least two studies, but FDA may not have
1549 critically assessed effect size and may not have closely analyzed “failed” studies. In
1550 general, FDA will usually not have carried out a meta-analysis of the trials. It is therefore
1551 essential to use the data from all available controlled trials (unless a trial has a significant
1552 defect), including trials conducted after marketing, to calculate a reasonable estimate of
1553 the actual control effect size, as described above. If the active-control data have been
1554 used to define a NI margin for another study, it is important to determine that the
1555 previous conclusion is applicable to the new study, but in general such prior use should
1556 indicate that FDA has assessed the NI margin for a NI study with similar endpoints and
1557 population.
1558

1559 **6. What are the choices of endpoints to be aware of before designing a non-inferiority**
1560 **trial design?**

1561
1562 The endpoints chosen for clinical trials (superiority or NI) reflect the event rate in the
1563 population, the importance of the event, and practical considerations, notably whether the
1564 event rates will allow a study of reasonable size. In NI studies, the endpoint must be one
1565 for which there is a good basis for knowing the effect of the active control. The endpoint
1566 used need not necessarily be the endpoint used in the historical trials or the effectiveness
1567 endpoint claimed in labeling. Past trials, for example, with mortality endpoints could, if
1568 data were available, be the basis for estimating an effect on a composite endpoint
1569 (cardiovascular mortality, myocardial infarction, and stroke), if that were the desired
1570 endpoint for the NI study. Such a change might be sought because it would permit a
1571 smaller study or was more feasible given current event rates.

1572

1573 **7. Are there circumstances where it may not be feasible to perform an NI study?**

1574

1575 Unfortunately, these are many, including some where a placebo-controlled study would
1576 not be considered ethical. Some examples include the following:

1577

- 1578 • The treatment effect may be so small that the sample size required to do a non-
1579 inferiority study may not be feasible.
- 1580 • There is large study-to-study variability in the treatment effect. In this case, the
1581 treatment effect may not be sufficiently reproducible to allow for the determination of
1582 a sufficiently reliable estimate of M_1 .
- 1583 • There is no historical evidence to determine a non-inferiority margin.
- 1584 • Medical practice has changed so much (e.g., the active control is always used with
1585 additional drugs) that the effect of the active control in the historical studies is not
1586 clearly relevant to the current study.

1587

1588 **8. In a situation where a placebo-controlled trial would be considered unethical, but a**
1589 **non-inferiority study cannot be performed, what are the options?**

1590

1591 In that case it may be possible to design a superiority study that would be considered
1592 ethical. These possibilities are discussed in section III of this guidance and ICH E-10,
1593 and include the following:

1594

- 1595 • When the new drug and established treatment are pharmacologically distinct, an add-
1596 on study where the test drug and placebo are each added to the established treatment.
- 1597 • A study in patients who do not respond to the established therapy. It may be possible
1598 to do a placebo-controlled trial in those patients. To establish specific effectiveness
1599 in non-responders, the study should randomize to test drug and the failed therapy and
1600 show superiority of the test drug.
- 1601 • A study in patients who cannot tolerate the established effective therapy.
- 1602 • A study of a population in which the effect of available therapy is not established.
- 1603 • For a drug with dose-related side effects, and where a dose lower than the usual dose
1604 would be considered ethical, a dose-response study may be possible.

1605

1606 **9. When will a single NI study be sufficient to support effectiveness?**

1607

1608 Several sections above touch on this question, notably III.B.2, which discusses it in
1609 detail. Briefly, reliance on a single study in the NI setting is based on considerations
1610 similar to reliance on a single study in the superiority setting, with the additional
1611 consideration of the stringency of showing NI using the M_2 NI margin. Many of these
1612 factors are described in the guidance for industry on *Providing Clinical Evidence of*
1613 *Effectiveness for Human Drugs and Biological Products*, and include prior supportive
1614 information, such as results with pharmacologically similar agents (a very common
1615 consideration, as the NI study will often compare drugs of the same pharmacologic
1616 class), support from credible biomarker information (tumor responses, ACE inhibition,

Contains Nonbinding Recommendations

Draft – Not for Implementation

1617 beta blockade), and a statistically persuasive result. With respect to the latter, it is noted
1618 above that a finding of NI based on excluding a treatment difference $> M_2$ provides very
1619 strong evidence (generally equivalent to a $p < 0.001$ in a superiority setting) that the test
1620 treatment has an effect > 0 . For all these reasons, most NI studies with outcome
1621 endpoints, if clearly successful, will be supportive as single studies. Of course, the
1622 importance of the study endpoint will influence the level of assurance needed, in a single
1623 study or multiple studies, that no more than M_2 has been lost.

1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668

APPENDIX — EXAMPLES

The following five examples derived from publicly available information (see references following examples) illustrate different aspects of the process of choosing a NI margin, of the application of a method of NI analysis, and other considerations relevant to whether it is possible to conduct and interpret the results of a NI study

Example 1(A): Determination of an NI Margin for a New Anticoagulant — Fixed Margin Approach

This example will demonstrate the following points:

- The determination of the NI margin (M_1) using the fixed margin approach
- How to select and assess the randomized trials of the active control on which to base the estimate of active comparator treatment effect.
- How to assess whether the assumption of assay sensitivity is appropriate, and whether the constancy assumption is reasonable for this drug class.
- Why it is appropriate to use a conservative choice (e.g., 95% lower bound) for estimating the treatment effect size of the active comparator, accounting for between-study variability, and considering other uncertainties in the randomized trial data.
- The use of the lower bound of the 95% confidence interval in the NI study for C-T to demonstrate non-inferiority.

SPORTIF V is an NI study that tested the novel anticoagulant ximelagatran against the active control warfarin. Warfarin is a highly effective, orally active anticoagulant that is approved in the United States for the treatment of patients with non-valvular atrial fibrillation at risk of thromboembolic complications (e.g., stroke, TIA, etc.). There are six placebo-controlled studies of warfarin involving the treatment of patients with non-valvular atrial fibrillation, all published between the years 1989 and 1993. The primary results of these studies are summarized in Table 1 and provide the basis for choosing the NI margin for SPORTIF V.

The point estimate of the event rate on warfarin compared to placebo is favorable to warfarin in each of the 6 studies. The upper bound of the 95% confidence interval of the risk ratio calculated in each study is less than one in five of the six studies, indicating a statistically demonstrated treatment effect in each of these studies. The one exception is the CAFA study. However, this study was reportedly stopped early because of favorable results published from the AFASAK and SPAF I studies (Connolly et al. 1991). Although the CAFA study was stopped early, a step that can sometimes lead to an overestimate of effect, the data from this study appear relevant in characterizing the overall evidence of effectiveness of warfarin because there is no reason to think it was stopped for early success, introducing a possible favorable bias. These placebo controlled studies of warfarin in

Contains Nonbinding Recommendations

Draft – Not for Implementation

1669 patients with non-valvular atrial fibrillation show a fairly consistent and reproducible effect.
1670 Based on the consistent results from the six studies, it can reasonably be assumed that were
1671 placebo to be included in a warfarin-controlled NI study involving a novel anticoagulant,
1672 warfarin would have been superior to placebo.

1673
1674

Table 1: Placebo-Controlled Trials of Warfarin in Non-Valvular Atrial Fibrillation

Study	Summary	Events/Patient Years		Risk Ratio (95% CI)
		Warfarin	Placebo	
AFASAK	open label. 1.2 yr follow-up	9/413 = 2.18%	21/398 = 5.28%	0.41 (0.19, 0.89)
BAATAF	open label. 2.2 yr follow-up	3/487 = 0.62%	13/435 = 2.99%	0.21 (0.06, 0.72)
EAFIT	open label. 2.3 yr follow-up patients with recent TIA	21/507 = 4.14%	54/405 = 13.3%	0.31 (0.19, 0.51)
CAFA*	double blind. 1.3 yr follow-up	7/237 = 2.95%	11/241 = 4.56%	0.65 (0.26, 1.64)
SPAF I	open label. 1.3 yr follow-up	8/260 = 3.08%	20/244 = 8.20%	0.38 (0.17, 0.84)
SPINAF	double blind. 1.7 yr follow-up	9/489 = 1.84%	24/483 = 4.97%	0.37 (0.17, 0.79)

1675

* CAFA was stopped early because of favorable results observed in other studies.

1676
1677
1678
1679
1680
1681
1682
1683
1684

As can be seen from the summary table, most of these studies were open label. It is not clear how great a concern this should be given the reasonably objective endpoints in the study (see Table 2), but to the extent there is judgment involved, there is some possible bias. The event rate on placebo in the EAFIT study was strikingly high, perhaps because the patient population in that study was different from the patient population studied in the remaining five studies in that only patients with a recent TIA or stroke were enrolled in EAFIT. That would clearly increase the event rate, but in fact the risk reduction in EAFIT was very similar to the four trials other than CAFA, which is relatively reassuring with respect to constancy of risk reduction in various AF populations.

1685
1686
1687
1688
1689
1690
1691
1692

Even if the historical studies are consistent, a critical consideration in deciding upon the NI margin derived from these studies is whether the constancy assumption is reasonable. The constancy assumption must consider whether the magnitude of effect of warfarin relative to placebo in the previous studies would be present in the new NI study, or whether changes in medical practice (e.g., concomitant medications, skill at reaching desired INR), or changes in the population being tested may make the effect of warfarin estimated from the previous studies not relevant to the current NI study.

1693
1694
1695
1696
1697
1698
1699
1700
1701

To evaluate the plausibility of this constancy assumption, one might compare some features of the six placebo-controlled warfarin studies with the NI study, SPORTIF V. There is considerable heterogeneity in the demographic characteristics of these studies. While some study subject characteristics can be compared across the studies (e.g., age, race, and target INR) certain characteristics cannot be compared (e.g., concomitant medication use, race, mean blood pressure at baseline) if they are not consistently reported in the study publications. Whether these are critical to outcomes is, of course, the critical question. Table 2 indicates that for some characteristics, such as a history of stroke or TIA, there are inter-study differences. One of the important inclusion criteria in the EAFIT study was that

Contains Nonbinding Recommendations

Draft – Not for Implementation

1702 subjects had a prior history of stroke or TIA. None of the other studies had such a
 1703 requirement. Subjects enrolled into the EAFT study were thus at higher risk than subjects in
 1704 the other studies, presumably leading to the higher event rates in both the warfarin and
 1705 placebo arms, shown in Table 1. The higher event rates in the EAFT study may also have
 1706 been influenced by the relatively long duration of follow-up or the fact that the primary
 1707 endpoint definition was broader, including vascular deaths and non-fatal myocardial
 1708 infarctions, which might have been less affected by coumadin, leading to a lower risk
 1709 reduction. This was not in fact seen. All in all, the results are quite consistent (with the
 1710 exception of CAFA), a relatively reassuring outcome.

1711

1712

Table 2: Demographic Variables, Clinical Characteristics, and Endpoints of Warfarin AF Studies

	AFASAK	BAATAF	CAFA	SPAF	VA	EAFT	SPORTIF V
Age years (mean)	73	69	68	65	67	71	72
Sex (%) Male	53%	75%	76%	74%	100%	59%	70%
h/o stroke or TIA (%)	6%	3%	3%	8%	0%	100%	18.3%
h/o HTN (%)	32%	51%	43%	49%	55%	43%	81%
≥65 years old & CAD (%)*	8%	10-16%	12-15%	7%	17%	7%	41%
>65 years old & DM (%)*	7-10%	14-16%	10-14%	13%	17%	12%	19%
h/o LV dysfunction (%)*	50%	24-28%	20-23%	9%	31%	8%	39%
Mean BP at BL (mm Hg)	NA	NA	NA	130/78	NA	145/84	133/77
Target INR	2.8-4.2	1.5-2.7	2-3	2-4.5	1.4-2.8	2.5-4.0	2-3
Primary endpoint	Stroke, TIA, systemic embolism	Ischemic stroke	Ischemic stroke and systemic embolism	Ischemic stroke and systemic embolism	Ischemic stroke	Vascular death, NF MI, stroke, systemic embolism	Stroke (ischemic + hemorrhagic) and systemic embolism

1713

1714

1715

1716

* = Not possible to verify whether definitions of CAD, DM, and LV dysfunction were the same in comparing the historic studies and SPORTIF V.
 NA = Not available

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

At the time the SPORTIF V study was reviewed, concerns about whether the constancy assumption held and other factors led to the consideration of whether discounting of the effect size would be appropriate (see discussion of discounting in section IV of this guidance). We now believe the historic results are reasonably likely to be consistent with results that would be seen today so that discounting was not necessary. To calculate M_1 , the relative risks in each of the six studies were combined using a random effects model to give a point estimate of 0.361 for the relative risk with a confidence interval of (0.248, 0.527). The 95% CI upper bound of 0.527 represents a 47% risk reduction, which translates into a risk increase of about 90% from not being on warfarin ($1/0.527 = 1.898$) (i.e., what would be seen if the test drug had no effect). Thus, M_1 (in terms of the hazard ratio favoring the control to be ruled out) is 1.898.

Contains Nonbinding Recommendations

Draft – Not for Implementation

1728

1729 It was considered clinically necessary to show that the test drug preserved a substantial
1730 fraction of the warfarin effect. The clinical margin M_2 representing the largest acceptable
1731 inferiority of the test to control, was therefore set at 50% of M_1 . As described in section IV
1732 of the guidance, we calculate M_2 , using the log hazard risk ratios, as 1.378, 95% CI for C-T <
1733 1.378.

1734

1735 In the SPORTIF V study, the point estimate of the relative risk was 1.39 and the two-sided
1736 95% confidence interval for the relative risk was (0.91, 2.12). Thus, in this example, the
1737 non-inferiority of ximelegatran to warfarin is not demonstrated because the upper limit (2.12)
1738 is greater than M_2 (=1.378). Indeed, it does not even demonstrate that M_1 (=1.898) has been
1739 excluded.

1740

1741 This example illustrates the fixed margin approach and what is often called the “two 95%
1742 confidence interval approach.” That is, a two-sided 95% confidence interval is used for the
1743 historical data to select M_1 , and a two-sided 95% confidence interval is used to test whether
1744 M_2 has been ruled out, similar to controlling the Type 1 error of the NI study at one-sided
1745 2.5%.

1746 **Example 1(B): Application of the Synthesis Method to the Above Example 1(A)**

1747

1748 This example demonstrates the following:

1749

1750 • The critical features of the synthesis approach to demonstrating the NI of a new
1751 anticoagulant.

1752

1753 • The calculations and sources of statistical variability that are incorporated in the
1754 synthesis approach.

1755

1756 • The main differences in interpretation of the fixed margin and the synthesis approaches
1757 when applied to the same set of studies and data.

1758

1759 In this example, we illustrate the synthesis method using the same data as Example 1(A),
1760 which consist of six studies comparing warfarin to placebo and one NI study comparing
1761 ximelegatran to warfarin. In contrast to the fixed margin method in Example 1(A), the
1762 synthesis method does not use a separate 95% confidence interval for this historical estimate
1763 of the effect of warfarin versus placebo and for the comparison in the NI study. Rather, the
1764 synthesis method is constructed to address the questions of whether ximelegatran preserves a
1765 specified percent, in this case 50% or one-half (versus placebo), of the effect of warfarin, and
1766 whether ximelegatran would be superior to a placebo, if one had been included as a
1767 randomized treatment group in the NI study. To accomplish this goal, the synthesis method
1768 makes a comparison of the effect of ximelegatran in the NI study to historical placebo data,
1769 an indirect comparison that is not based upon a randomized current placebo group. The
1770 synthesis method combines the data from the placebo-controlled studies of warfarin with the
1771 data from the NI study in such a way that a test of hypothesis is made to demonstrate that a
1772 certain percent of the effect of warfarin is retained in the NI study. A critical point
1773 distinguishing the synthesis method from the fixed margin method is that the M_1 effect size
1774 of warfarin is not specified in advance and is not required to be fixed prior to carrying out the
1775 synthesis method. But to carry out the analysis, an assumption needs to be made regarding
1776 the placebo comparison, namely, that the difference between control drug and placebo (had
1777 there been one) in the NI trial is the same as what was seen in the historical placebo-
1778 controlled trials of warfarin. The assumption is needed because there is no randomized
1779 comparison of warfarin and placebo in the NI trial. As a point of reference, we know from
1780 the previous example, 1(A), that the warfarin effect M_1 was estimated from the historical
1781 placebo studies to be a 47% risk reduction.

1782

1783 In this case, the synthesis method statistically tests the null hypothesis that the inferiority of
1784 ximelegatran compared to warfarin is less than 50% or one half of the risk reduction of
1785 warfarin compared to placebo, a question that the fixed margin method does not directly
1786 address because in the fixed margin method, the placebo is only present in the historical
1787 studies and not in the NI study. We carry out this test on the log relative risk scale, so that
1788 the null hypothesis can be written as:

Contains Nonbinding Recommendations

Draft – Not for Implementation

1789

1790

$H_0: \{\log\text{-Relative Risk of ximelegatran versus warfarin}\} \geq$

1791

$-\frac{1}{2} \{\log\text{-Mean Relative Risk of warfarin versus placebo}\}$

1792

A test of this hypothesis is performed by the expression below (the statistical test) that has the form of a quotient where the numerator is an estimate of the parameter defined in the null hypothesis by $\{\log\text{-Relative Risk of ximelegatran versus warfarin}\} + \frac{1}{2} \{\log\text{-Mean Relative Risk of warfarin versus placebo}\}$ and the denominator is an estimate of the standard error of the numerator. In this case, the estimated log-Relative Risk of ximelegatran versus warfarin is 0.329 (log of 1.39) with a standard error of 0.216 while the estimated log-Relative Risk of warfarin versus placebo is -1.02 (log of .527) with a standard error of 0.154. The estimate of the log warfarin effect is -1.02, and the standard error of this estimate is 0.154; these estimates are combined with the NI data as if all the data were in a randomized comparison with placebo. The synthesis test statistic is calculated as:

1793

1794

1795

1796

1797

1798

1799

1800

1801

1802

1803

$$\frac{0.329 + \frac{1}{2}\{-1.02\}}{\sqrt{0.216^2 + \left\{\frac{1}{2}\{0.154\}\right\}^2}} = -0.789$$

1804

1805

1806

1807

1808

1809

Assuming the statistic is normally distributed, it is then compared to -1.96 (for one-sided Type 1 error rate of 0.025). For this case, the value, -0.789, is not less (more negative) than -1.96, so we cannot reject the null hypothesis. Therefore, it cannot be concluded that an NI margin of 50% retention is satisfied.

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

To compare the fixed margin method with the synthesis method, recall that the fixed margin compares the upper or lower limits of two 95% confidence intervals, one for the NI study and one for the meta-analysis of the effect of warfarin. One might consider the fixed margin approach as conservative, as it compares to statistically “worst cases.” The synthesis method does not use two such worst cases. To provide a more detailed comparison of the approaches, the fixed margin approach can be expressed as using a test statistic similar to that of the synthesis approach.

The synthesis method concludes non-inferiority if

1820

$$\frac{0.329 + \frac{1}{2}\{-1.02\}}{\sqrt{0.216^2 + \left\{\frac{1}{2}\{0.154\}\right\}^2}} < -1.96$$

1821

Contains Nonbinding Recommendations

Draft – Not for Implementation

1822

1823 The fixed margin method concludes non-inferiority if

1824

1825
$$\frac{0.329 + \frac{1}{2}\{-1.02\}}{0.216 + \frac{1}{2}\{0.154\}} < -1.96$$

1826

1827 The critical difference between these two procedures is the form of the denominator, which
1828 expresses the standard errors of the expressions in the numerator. The synthesis standard
1829 error is always smaller than that of the fixed margin method when expressed in this manner.
1830 In most situations, the synthesis is therefore statistically more efficient (and would require a
1831 smaller sample size) than the fixed margin approach. Of course, the approach can be
1832 considered useful and valid only if the assumptions of the synthesis method can be
1833 considered satisfied. This is not always possible, generally because of concerns about
1834 constancy, that is, whether the historical differences from placebo would accurately describe
1835 the current differences from placebo.

1836

1837 The two procedures also cannot be directly compared because they have other differences
1838 that make their comparison problematic, notably the differences in how the statistical error
1839 rates, or Type 1 errors, are calculated and interpreted. The synthesis method, because of the
1840 way it makes the comparisons with a placebo, gives equal weight to the variance (or
1841 variability of the outcome data) in this historical estimate and the variance of the data
1842 obtained from the randomized comparison of the test drug and active comparator in the NI
1843 study. When the historical database is very large relative to the NI database, combining the
1844 historical data and NI together may suggest greater precision in the overall assessment of the
1845 NI study than is warranted given the fact that the placebo comparisons were from studies
1846 conducted in a different population, usually at a different time. In contrast, the fixed margin
1847 method controls a Type 1 error rate within the NI study that is conditioned on the pre-
1848 specified fixed NI margin, separately estimated from the historical active comparator data.
1849 The synthesis test method also does not estimate a fixed NI margin to be excluded (i.e., one
1850 depending only on the prior placebo-controlled data for the active comparator).

1851

1852 A general principle expressed in this guidance is the need to be conservative in the selection
1853 of the margin M_1 because that margin is critical to establishing that a test drug is effective in
1854 an NI study design. The M_1 margin is usually chosen conservatively because of the
1855 uncertainties associated with the validity of assumptions in an NI study and the reliance on
1856 historical active control comparisons. As noted, the fixed margin approach can be
1857 considered conservative in that several worst case situations (lower bounds of 95%
1858 confidence intervals) are used, one evaluating the historical evidence and another in the NI
1859 comparison. We recommend use of this conservative fixed margin approach to selecting the
1860 M_1 margin and to demonstrating in the NI study that the M_1 margin is excluded at the
1861 acceptable Type 1 error. The synthesis method, on the other hand, as described above, is less
1862 conservative. But this is reasonable, given that M_2 is considerably smaller (a more
1863 demanding margin) and that the presence of a control drug effect has been well established
1864 by ruling out loss of M_1 using the fixed margin approach. We therefore believe the NI study

Contains Nonbinding Recommendations

Draft – Not for Implementation

1865 should utilize a fixed margin approach to ruling out loss of M_1 but can use the synthesis
1866 method to establish that loss of effect greater than the clinically relevant margin M_2 has been
1867 ruled out.

1868 **Example 2: The Determination of a Non-Inferiority Margin for Complicated Urinary**
1869 **Tract Infection (cUTI) — Fixed Margin Approach**

1870

1871 This example will illustrate the following points:

1872

- 1873 • The use of the absolute difference in cure rates as the metric of treatment effect.
- 1874 • The determination of a non-inferiority margin when there are no randomized active
1875 comparator placebo-controlled studies available for the indication of interest (in this
1876 case, cUTI).
- 1877 • Estimating the placebo response rate in cUTI based upon data from uncomplicated
1878 urinary tract infections (a generally less severe form of urinary tract infection leading
1879 to a high, therefore conservative, estimate).
- 1880 • The importance of seeking out all relevant studies for the margin determination and
1881 incorporating the limitations of the studies, the analyses, and the resulting estimates in
1882 the consideration of the resulting estimate of the non-inferiority margin.
- 1883 • This approach (i.e., relying on data other than controlled trials of the active control) is
1884 credible only when the effect size is large, given its limitations.

1885

1886 The following steps were used to estimate the effectiveness of the active control.

1887

- 1888 1. Evaluation of the placebo response rate in uncomplicated urinary tract infection
1889 (uUTI)
- 1890 2. Evaluation of outcomes in patients receiving inadequate or inappropriate therapy for
1891 complicated urinary tract infection (cUTI)/acute pyelonephritis (AP)
- 1892 3. Evaluation of the active comparator's response rate (levofloxacin, in this case) for
1893 cUTI.

1894

1895 **Step 1: Placebo Response Rate for Uncomplicated Urinary Tract Infection (uUTI)**

1896

1897 Although there were no placebo-controlled complicated UTI studies available, three placebo-
1898 controlled studies in women with uncomplicated UTI were identified. Among these three
1899 studies there were differences in the duration of study drug, endpoints assessed, and the
1900 diagnostic criteria for significant bacteriuria. There were no placebo-controlled trials
1901 identified in men with UTI without significant co-morbid conditions, and the
1902 pathophysiology and natural history of UTI are different in men and women. It would be
1903 expected that placebo response rates would therefore be high in such studies compared to the
1904 untreated rate in cUTI and represent a conservative (high) estimate of the spontaneous cure
1905 rate in cUTI.

1906

1907 Microbiological eradication rate is generally used as the primary endpoint for UTI studies.
1908 In the three placebo-controlled studies identified for UTI, the bacteriological response rates
1909 were 95/227(42%) for the combined 8-10 and 35-49 days (Ferry et al.), 9/27(33%) at day 3
1910 (Christiaens et al.), and 8/18(44%) in 1 week (Dubi et al.). The bacteriologic criteria for
1911 entry used in the Ferry study were $\geq 10^3$ CFU/ml for primary pathogens, whereas $\geq 10^4$
1912 CFU/ml was used for the Christiaens study. Because a count of $\geq 10^5$ CFU/ml is more

1913 typically used as diagnostic criteria for a uropathogen, the studies could overestimate the
 1914 placebo response rates by including patients whose colony counts would not cause them to be
 1915 considered infected. The results are summarized in the following table.
 1916

Author	Type of UTI	Placebo	95% CI ¹
Ferry et al.	uUTI	95/227 (42%)	(35.4 %, 48.6%)
Christiaens et al.	Acute uUTI	9/27 (33%)	(16.5%, 54.0%)
Dubi et al.	uUTI	8/18 (44%)	(21.5%, 69.2%)

1917 ¹Exact Confidence Intervals

1918
 1919 Because of the unequal study population sizes, a weighted analysis is needed. The weighted
 1920 non-iterative method for random effects model using logit of the event rates described by
 1921 DerSimonian and Laird was used to obtain the estimate and its 95% CI; the weighted
 1922 estimate is 41.2% with 95% CI of (35.5%, 47.2%).
 1923

1924 **Step 2: Outcomes Subsequent to Inadequate or Inappropriate Antibacterial Therapy**
 1925 **for Complicated Urinary Tract Infection (cUTI)/AP**

1926
 1927 Three studies were identified in which some patients were treated with an antimicrobial drug
 1928 to which the bacteria causing their UTI were resistant (inadequate therapy). Eradication rates
 1929 for pathogens resistant to the antimicrobial drug may be considered as another way to
 1930 estimate the placebo effect in cUTI/AP. It should be noted, however, that the use of data
 1931 from inadequate therapy may result in an estimate that is higher than a true placebo, once
 1932 again a conservative estimate of effect, because even “inadequate” therapy may have some
 1933 effect on the patient’s infection.
 1934

Author	Type of UTI	Eradication Rates	95% CI ¹
Allais et al.	cUTI/AP	12/23 (52.2%)	(30.6%, 73.2%)
Fang et al.	cUTI/AP	4/28 (14.3%)	(4.0%, 32.7%)
Talan et al.	AP	7/14 (50.0%)	(23.0%, 77.0%)

1935 ¹Exact Confidence Intervals

1936
 1937 The data from the historical studies in Table 4 were combined to obtain a weighted estimate
 1938 of the inadequate therapy eradication rate and its corresponding two-sided 95% CI. The
 1939 weighted estimate using the DerSimonian and Laird approach (random effect model) is
 1940 36.8% with 95% CI of (15.4%, 64.9%).
 1941

1942 **Step 3: Active Comparator's Eradication Rate for Complicated UTI (cUTI)**

1943
 1944 To assess the eradication rates for the active comparator, levofloxacin, four cUTI studies
 1945 were considered, including two published studies and two studies submitted to the Agency
 1946 (Study A and Study B) that involved men and women ≥18 years old. The two studies from

1947 the medical literature had limitations. In the Peng study, the microbiological eradication rate
 1948 was evaluated on Day 5, while antibiotic therapy was still ongoing. This could have falsely
 1949 elevated the response rate. The Klimberg study was an open-label study, and was excluded
 1950 from the analysis because of concern about potential bias.

1951
 1952 The other two studies, Study A and Study B, were blinded controlled studies using
 1953 levofloxacin for the treatment of cUTI. In Study A, the microbiological eradication rate for
 1954 levofloxacin was 84.2% (154/183). In Study B, the microbiological eradication rate for
 1955 levofloxacin was 78.2% (252/321). The levofloxacin eradication rates for the Peng study and
 1956 Studies A and B are shown in Table 5. The weighted estimate of eradication rates using the
 1957 DerSimonian and Laird approach is 81.6% with 95% CI of (75.8%, 86.3%).
 1958

Author	Type of UTI	Levofloxacin Microbiological Eradication Rate	95% CI ¹
Peng et al.	cUTI	18/20 (90%)	(68.3%, 98.8%)
Study A	cUTI and AP	154/183 (84.2%)	(78.0%, 89.1%)
Study B	cUTI and AP	252/321 (78.2%)	(73.6%, 82.9%)

1959 ¹Exact confidence intervals
 1960

1961 **Step 4: Estimated Non-Inferiority Margin for Complicated UTI (cUTI) Using**
 1962 **Levofloxacin as the Active Comparator**
 1963

1964 The placebo eradication rate is estimated from the upper bound of the two-sided 95% CI for
 1965 the placebo eradication rate in uUTI (47%) and this estimate is supported by evidence based
 1966 on outcomes subsequent to inadequate or inappropriate therapy in cUTI (65%). The
 1967 estimated levofloxacin cure rate for sensitive organisms is 76% (using the lower bound of the
 1968 95% CI for the weighted levofloxacin response rate). Using the placebo eradication rate for
 1969 uUTI, the historical treatment effect can be calculated as 29% (=76%-47%). The treatment
 1970 effect based on outcomes following inadequate antibacterial therapy can be calculated as
 1971 11% (=76%-65%), providing supportive evidence.
 1972

1973 **Major Limitations in This Example:**
 1974

1975 Apart from the lack of a direct comparison of active control and placebo in cUTI, there were
 1976 various uncertainties in the historical estimates described above because of problems with
 1977 data quality, study design, population size, prognostic factors, and differences in the timing
 1978 of the microbiological endpoint assessments. On the other hand, the placebo eradication rate
 1979 was estimated based on placebo-controlled clinical studies assessing the antibacterial
 1980 treatment in a population (female subjects with uUTI) that would almost certainly give an
 1981 overestimate of the spontaneous or placebo eradication rate in cUTI, leading to a
 1982 conservative (low) estimate of the effect of the active control.

Contains Nonbinding Recommendations

Draft – Not for Implementation

1983

1984 **Discounting and Preservation of the Levofloxacin Treatment Effect:**

1985

1986 The various limitations and uncertainties in the historical data led to discounting of the
1987 calculated treatment effect of 29%. Thus, the active control treatment effect over placebo
1988 (M_1) was estimated as 14.5% based on a 50% discounting. For a serious illness, a substantial
1989 portion (at least 50% or more) of M_1 should be preserved. Accordingly, an NI margin of 7%
1990 was specified as M_2 based on clinical judgment.

1991

1992 **Example 3: Aspirin to Prevent Death or Death/MI After Myocardial Infarction**

1993

1994 This example demonstrates the following:

1995

- 1996 • When it may not be possible to determine the NI margin because of the limitations of the
- 1997 data available.

1998

1999 By 1993, the effect of aspirin in preventing death after myocardial infarction had been
 2000 studied in six large randomized placebo-controlled clinical trials. A seventh trial, ISIS-2,
 2001 gave the drug during the first day after the AMI and is not included because it addressed a
 2002 different question. The results are summarized and presented in chronological order in Table
 2003 6.

2004

2005 **Table 6. Results of six placebo-controlled randomized studies (listed in chronological order) of the effect**
 2006 **of aspirin in preventing death after myocardial infarction**

Study	Year published	Aspirin		Placebo		Relative Risk (95% CI)
		N	Death rate	N	Death rate	
MRC-1	1974	615	8.0%	624	10.7%	0.74 (0.52, 1.05)
CDP	1976	758	5.8%	771	8.3%	0.70 (0.48, 1.01)
MRC-2	1979	832	12.2%	850	14.8%	0.83 (0.65, 1.05)
GASP	1978	317	10.1%	309	12.3%	0.82 (0.53, 1.28)
PARIS	1980	810	10.5%	406	12.8%	0.82 (0.59, 1.13)
AMIS	1980	2267	10.9%	2257	9.7%	1.12 (0.94, 1.33)

2007

2008 The results suggest:

2009

- 2010 (1) The effect of aspirin on mortality as measured by the relative risk seems to
- 2011 attenuate over the time the studies were conducted.
- 2012 (2) The largest trial, AMIS, showed a numerically adverse effect of aspirin.

2013

2014 The relative risk in the AMIS study is significantly different from the mean relative risk in
 2015 the remaining studies ($p \leq 0.005$). The validity of pooling the results of AMIS with those of
 2016 the remaining studies is therefore a concern. It would be invalid to exclude AMIS from the
 2017 meta-analyses because its effect differed from the effect in the remaining studies, unless there
 2018 were adequate clinical or scientific reasons for such exclusion. At a minimum, any meta-
 2019 analysis of all studies would need to reflect this heterogeneity by using a random-effect
 2020 analysis.

2021

2022 Although a fixed effect analysis of the six studies gives a point estimate of 0.91 (95% CI 0.82
 2023 to 1.02), the random-effects analysis gives a point estimate of 0.86 with 95% confidence
 2024 interval (0.69, 1.08). The effect of aspirin on prevention of death after myocardial infarction
 2025 in these historical studies is thus inconclusive (i.e., the upper bound of the 95% CI for effect
 2026 is > 1.0). Therefore, it would be difficult, indeed not really possible, to select aspirin as the

Contains Nonbinding Recommendations

Draft – Not for Implementation

2027 active control for evaluating the mortality effect of a test drug in a non-inferiority trial. Apart
 2028 from this calculation, it seems difficult to accept an NI endpoint that is not supported by the
 2029 largest of the six trials.

2030

2031 The same six studies can also be examined for the combined endpoint of death plus AMI in
 2032 patients with recent AMI. This endpoint reflects the current physician-directed claim for
 2033 aspirin based on the positive finding in two studies (MRC-2, PARIS).

2034

2035 **Table 7. Results of six placebo-controlled randomized studies of the effect of aspirin in secondary**
 2036 **prevention of death or MI after myocardial infarction**

Study	Year published	Aspirin		Placebo		Relative Risk (95% CI)
		N	Event rate	N	Event rate	
MRC-1	1974	615	9.9%	624	13.1%	0.75 (0.55, 1.03)
CDP	1976	758	9.5%	771	12.5%	0.76 (0.57, 1.02)
MRC-2	1979	832	16.0%	850	22.2%	0.72 (0.59, 0.88)
GASP	1978	317	13.6%	309	17.5%	0.78 (0.54, 1.12)
PARIS	1980	810	17.4%	406	22.7%	0.77 (0.61, 0.97)
AMIS	1980	2267	18.6%	2257	19.2%	0.97 (0.86, 1.09)

2037

***the event rate of either group needs further verification from each article**

2038

2039 The results indicate that the effect of aspirin on death or MI after myocardial infarction is
 2040 small to absent in the latest trial (AMIS). Random-effect analyses give, depending on the
 2041 specific analysis, point estimates of the relative risk of 0.81-0.85, with 95% CI upper bounds
 2042 of 0.96-1.02. The NI margin based on these six studies ranges from 4% to zero (without
 2043 reducing it further to represent M_2) is so small that a trial to rule out loss at this effect would
 2044 be unrealistically large. Again, as with the mortality endpoint, it would be troubling even to
 2045 consider an NI approach when the largest and most recent trial showed no significant effect.

2046 **Example 4: Xeloda to Treat Metastatic Colorectal Cancer - the Synthesis Method**

2047

2048 This example of Xeloda for first-line treatment of metastatic colorectal cancer illustrates:

2049

- 2050 • The use of the synthesis method to demonstrate a loss of no more than 50% of the
- 2051 historical control treatment’s effect and a relaxation of this criterion when two NI studies
- 2052 are available.
- 2053
- 2054 • The use of supportive endpoints in the decision making process.
- 2055
- 2056 • The use of a conservative estimate of the control treatment effect size, because a subset
- 2057 of the available studies to estimate the margin was selected and the effect was measured
- 2058 relative to a previous standard of care instead of placebo.
- 2059

2060 The U.S. regulatory standard for first-line treatment of metastatic colorectal cancer, the use

2061 sought for Xeloda, is the demonstration of improvement in overall survival. Two separate

2062 clinical trials, each using an NI study design, compared Xeloda to a Mayo Clinic regimen of

2063 5-fluorouracil with leucovorin (5-FU+LV), the standard of care at the time. Xeloda is an oral

2064 fluoropyrimidine, while 5-fluorouracil (5-FU) is an infusional fluoropyrimidine

2065

2066 By itself, bolus 5-FU had not demonstrated a survival advantage in first-line metastatic

2067 colorectal cancer. But with the addition of leucovorin to bolus 5-FU, the combination had

2068 demonstrated improved survival. A systematic evaluation of approximately 30 studies that

2069 investigated the effect of adding leucovorin to a regimen of 5-FU identified ten clinical trials

2070 that compared a regimen of 5-FU+LV similar to the Mayo clinic regimen to 5-FU alone,

2071 thereby providing a measure of the effect of LV added to 5-FU, a conservative estimate of

2072 the overall effect of 5-FU+LV, as it is likely 5-FU has some effect.

2073

2074 Table 8 summarizes the overall survival results, using the metric “log hazard ratio” for the

2075 ten studies identified that addressed the comparison of interest.

2076

2077 **Table 8: Selected studies comparing 5FU to 5-FU+LV**

Study	Hazard Ratio ¹	Log Hazard Ratio ¹	Standard Error
Historical Study 1	1.35	.301	.232
Historical Study 2	1.26	.235	.188
Historical Study 3	0.78	-.253	.171
Historical Study 4	1.15	.143	.153
Historical Study 5	1.39	.329	.185
Historical Study 6	1.35	.300	.184
Historical Study 7	1.38	.324	.166
Historical Study 8	1.34	.294	.126
Historical Study 9	1.03	.0296	.165
Historical Study 10	1.95	.670	.172

2078 ¹ All log hazard ratios are 5-FU/5-FU+LV

Contains Nonbinding Recommendations

Draft – Not for Implementation

2079 A random effects model applied to the survival results of these ten studies yielded the
2080 historical estimate of the 5-FU versus 5-FU+LV survival comparison of log hazard ratio of
2081 1.264 with a 95% confidence interval of (1.09, 1.46) and a log hazard ratio of 0.234. The NI
2082 margin is therefore 1.09 for a fixed margin approach ruling out M_1 .
2083

2084 A summary of the survival results based on the intent-to-treat populations for each of the two
2085 Xeloda NI trials is presented in Table 9. Study 2 rules out M_1 using a fixed margin approach,
2086 but Study 1 does not.
2087

2088 **Table 9: Summary of the survival results**

Study	Hazard Ratio ¹	Log Hazard Ratio ¹	Standard Error	95% CI for the Hazard Ratio ¹
NI Study 1	1.00	-0.0036	0.0868	(0.84, 1.18)
NI Study 2	0.92	-0.0844	0.0867	(0.78, 1.09)

2089 ¹ Hazard ratios and log hazard ratios are Xeloda/5-FU+LV
2090

2091 The clinical choice of how much of the effect on survival of 5-FU+LV should be shown not
2092 to be lost by Xeloda was determined to be 50%. The synthesis approach was used to analyze
2093 whether the NI criteria of 50% loss was met. This synthesis approach to the non-inferiority
2094 test procedure for each study combines the results of each NI study with the results from the
2095 random effects meta-analysis into a normalized test statistic.
2096

2097 Based on this NI synthesis test procedure, NI Study 1 failed to demonstrate that Xeloda
2098 retained at least 50% of the historical effect of 5-FU+LV versus 5-FU on overall survival, but
2099 NI study 2 did demonstrate such an effect. It was then decided to determine what percent
2100 retention might be satisfied by the data in a statistically persuasive way. By adapting the
2101 synthesis test procedure for retention of an arbitrary percent of the 5-FU+LV historical effect,
2102 it was determined that NI Study 1 demonstrated that Xeloda lost no more than 90% of the
2103 historical effect of 5-FU+LV on overall survival and that NI Study 2 demonstrated no more
2104 than a 39% loss of the historical effect.
2105

2106 The evidence of effectiveness of Xeloda was supported by the observation that the tumor
2107 response rates were statistically significantly greater for the Xeloda arm and the fact that
2108 Xeloda and 5-FU were structurally and pharmacologically very similar.
2109

2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154

REFERENCES - EXAMPLES

Example 1(A)

The Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators (1990). “The Effect of Low-Dose Warfarin on the Risk of Stroke in Patients with Nonrheumatic Atrial Fibrillation.” *New Engl J Med* 323, 1505-1511.

Connolly, S.J., Laupacis, A., Gent, M., Roberts, R.S., Cairns, J.A., Joyner, C. (1991). “Canadian Atrial Fibrillation Anticoagulation (CAFA) Study.” *J Am Coll Cardiol* 18, 349-355.

EAFT (European Atrial Fibrillation Trial) Study Group (1993). “Secondary Prevention in Non-Rheumatic Atrial Fibrillation After Transient Ischemic Attack or Minor Stroke.” *Lancet* 342, 1255-1262.

Ezekowitz, M.D., Bridgers, S.L., James, K.E., Carliner, N.H., et al. (1992). “Warfarin in the Prevention of Stroke Associated with Nonrheumatic Atrial Fibrillation.” *N Engl J Med* 327, 1406-1412.

Food and Drug Administration, Dockets home page. Available at: http://www.fda.gov/ohrms/dockets/AC/04/briefing/2004-4069B1_07_FDA-Backgrounder-C-R-stat%20Review.pdf.

Halperin, J.L., Executive Steering Committee, SPORTIF III and V Study Investigators (2003). “Ximelagatran Compared with Warfarin for Prevention of Thromboembolism in Patients with Nonvalvular Atrial Fibrillation: Rationale, Objectives, and Design of a Pair of Clinical Studies and Baseline Patient Characteristics (SPORTIF III and V).” *Am Heart J* 146, 431-8.

Jackson, K., Gersh, B.J., Stockbridge, N., Fleiming, T.R., Temple, R., Califf, R.M., Connolly, S.J., Wallentin, L., Granger, C.B. (2005). Participants in the Duke Clinical Research Institute/American Heart Journal Expert Meeting on Antithrombotic Drug Development for Atrial Fibrillation (2008). “Antithrombotic Drug Development for Atrial Fibrillation: Proceedings.” Washington, D.C., July 25-27, 2005. *American Heart Journal* 155, 829-839.

Petersen, P., Boysen, G., Godtfredsen, J., Andersen, E.D., Andersen, B. (1989). “Placebo-Controlled, Randomised Trial of Warfarin and Aspirin for Prevention of Thromboembolic Complications in Chronic Atrial Fibrillation.” *The Lancet* 338, 175-179.

Stroke Prevention in Atrial Fibrillation Investigators (1991). “Stroke Prevention in Atrial Fibrillation Study: Final Results.” *Circulation* 84, 527-539.

Contains Nonbinding Recommendations

Draft – Not for Implementation

2155

2156 **Example 1(B) Refer to "General Reference" Section for synthesis methods.**

2157

2158 **Example 2**

2159

2160 Allais, J.M., Preheim, L.C., Cuevas, T.A., Roccaforte, J.S., Mellencamp, M.A., Bittner, M.J.
2161 (1988). "Randomized, Double-Blind Comparison of Ciprofloxacin and Trimethoprim
2162 Sulfamethoxazole for Complicated Urinary Tract Infections." *Antimicrob Agents*
2163 *Chemother.* 32(9), 1327-30.

2164

2165 Christiaens, T.C., De Meyere, M., Verschraegen, G., et al (2002). "Randomised Controlled
2166 Trial of Nitrofurantoin Versus Placebo in the Treatment of Uncomplicated Urinary Tract
2167 Infection in Adult Women." *Br J Gen Pract.* 52(482), 729-34.

2168

2169 DerSimonian, R., Laird, N. (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical*
2170 *Trials.* 7, 177-188.

2171

2172 Dubi, J., Chappuis, P., Darioli, R. (1982). "Treatment of Urinary Infection with a Single
2173 Dose of Co-trimoxazole Compared with a Single Dose of Amoxicillin and a Placebo."
2174 *Schweiz Med Wochenschr.* 12(3), 90-92.

2175

2176 Fang, G.D., Brennen, C., Wagener, M. et al (1991). "Use of Ciprofloxacin Versus Use of
2177 Aminoglycosides for Therapy of Complicated Urinary Tract Infection: Prospective,
2178 Randomized Clinical and Pharmacokinetic Study." *Antimicrob Agents Chemother.* 35(9),
2179 1849-55.

2180

2181 Ferry, S.A., Holm, S.E., Stenlund, H., Lundholm, R., Monsen, T.J. (2004). "The Natural
2182 Course of Uncomplicated Lower Urinary Tract Infection in Women Illustrated by a
2183 Randomized Placebo-Controlled Study." *Scan J Infect Dis.* 36, 296-301.

2184

2185 Ferry, S.A., Holm, S.E., Stenlund, H., Lundholm, R., Monsen, T.J. (2007). "Clinical and
2186 Bacteriological Outcome of Different Doses and Duration of Pivmecillinam Compared with
2187 Placebo Therapy of Uncomplicated Lower Urinary Tract Infection in Women: The LUTIW
2188 Project." *Scan J of Primary Health Care.* 25(1), 49-57.

2189

2190 Klimberg, I.W., Cox, C.E. 2nd, Fowler, C.L., King, W., Kim, S.S., Callery-D'Amico, S.
2191 (1998). "A Controlled Trial of Levofloxacin and Lomefloxacin in the Treatment of
2192 Complicated Urinary Tract Infection." *Urology.* 51(4), 610-5.

2193

2194 Peng, M.Y. (1999). "Randomized, Double-Blind, Comparative Study of Levofloxacin and
2195 Ofloxacin in the Treatment of Complicated Urinary Tract Infections." *J Microbiol Immunol*
2196 *Infect.* 32(1), 33-9.

2197

Contains Nonbinding Recommendations

Draft – Not for Implementation

2198 Talan, D.A., Stamm, W.E., Hooton, T.M. et al (2000). “Comparison of Ciprofloxacin (7
2199 Days) and Trimethoprim-Sulfamethoxazole (14 Days) for Acute Uncomplicated
2200 Pyelonephritis in Women.” *JAMA*. 283(12), 1583-1590.

2201

2202 **Example 3**

2203

2204 Aspirin Myocardial Infarction Study Research Group (1980). “A Randomized Controlled
2205 Trial of Aspirin in Persons Recovered from Myocardial Infarction.” *JAMA* 243, 661-669.

2206

2207 Breddin, K., Loew, D., Lechner, K., Uberia, E.W. (1979). “Secondary Prevention of
2208 Myocardial Infarction. Comparison of Acetylsalicylic Acid, Phenprocoumon and Placebo. A
2209 Multicenter Two-Year Prospective Study.” *Thrombosis and Haemostasis* 41, 225-236.

2210

2211 Coronary Drug Project Group (1976). “Aspirin in Coronary Heart Disease.” *Journal of*
2212 *Chronic Disease* 29, 625-642.

2213

2214 Elwood, P.C., Cochrane, A.L., Burr, M.L., Sweetnam, P.M., Williams, G., Welsby, E.,
2215 Hughes, S.J., Renton, R. (1974). “A Randomized Controlled Trial of Acetyl Salicylic Acid
2216 in the Secondary Prevention of Mortality from Myocardial Infarction.” *British Medical*
2217 *Journal* 1, 436-440.

2218

2219 Elwood, P.C., Sweetnam, P.M. (1979). “Aspirin and Secondary Mortality After Myocardial
2220 Infarction.” *Lancet* ii, 1313-1215.

2221

2222 Fleiss, J.L. (1993). “The Statistical Basis of Meta-Analysis.” *Statistical Methods in Medical*
2223 *Research* 2,121-145.

2224

2225 ISIS-2 Collaborative Group (1988). “Randomized Trial of Intravenous Streptokinase, Oral
2226 Aspirin, Both, or Neither Among 17187 Cases of Suspected Acute Myocardial Infarction:
2227 ISIS-2.” *Lancet* 2, 349-360.

2228

2229 Persantine-Aspirin Reinfarction Study Research Group (1980). “Persantine and Aspirin in
2230 Coronary Heart Disease.” *Circulation* 62, 449-461.

2231

2232 **Example 4**

2233

2234 FDA Guidance for Industry: Oncologic Drugs Advisory Committee Discussion on FDA
2235 Requirements for the Approval of New Drugs for Treatment of Colon and Rectal Cancer.

2236

2237 FDA Medical-Statistical review for Xeloda (NDA 20-896) dated April 23, 2001.

2238 (http://www.fda.gov/cder/foi/nda/2001/20896s6_Xeloda_Medr_Statr_P1.pdf).

2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283

GENERAL REFERENCES

Blackwelder, W.C. (1982). “Proving the Null Hypothesis in Clinical Trials.” *Controlled Clinical Trials* 3, 345-353.

Blackwelder, W.C. (2002). “Showing a Treatment is Good Because it is Not Bad: When Does “Noninferiority” Imply Effectiveness?” *Control Clinical Trials* 23, 52–54.

Brittain, E., Lin, D. (2005). “A Comparison of Intent-to-Treat and Per Protocol Results in Antibiotic Non-Inferiority Trials.” *Statistics in Medicine* 24, 1-10.

Brown, D., Day, S. (2007). Reply. *Statistics in Medicine* 26, 234-236.

CBER/FDA Memorandum (1999). Summary of CBER Considerations on Selected Aspects of Active Controlled Trial Design and Analysis for the Evaluation of Thrombolytics in Acute MI, June 1999.

Committee for Proprietary Medicinal Products (CPMP) (2000). Points to Consider on Switching Between Superiority and Non-Inferiority. <http://www.emea.europa.eu/pdfs/human/ewp/048299en.pdf>.

Committee for Medicinal Products for Human Use (CHMP) (2006). “Guideline on the Choice of the Non-Inferiority Margin.” *Statistics in Medicine* 25, 1628–1638.

Chow, S.C., Shao, J. (2006). “On Non-Inferiority Margin and Statistical Tests in Active Control Trial.” *Statistics in Medicine* 25, 1101–1113.

D’Agostino, R.B., Massaro, J.M., Sullivan, L. (2003). “Non-Inferiority Trials: Design Concepts and Issues – the Encounters of Academic Consultants in Statistics.” *Statistics in Medicine* 22, 169–186.

D’Agostino, R.B., Campbell, M., Greenhouse, J. (2005). “Non-Inferiority Trials: Continued Advancements in Concepts and Methodology.” *Statistics in Medicine* 25, 1097-1099.

DerSimonian, R., Laird, N. (1986). “Meta-Analysis in Clinical Trials.” *Controlled Clinical Trials* 7, 177-188.

Ellenberg, S.S., Temple, R. (2000). “Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments - Part 2: Practical Issues and Specific Cases.” *Annals of Internal Medicine* 133, 464-470.

Fisher, L.D., Gent, M., Büller, H.R. (2001). “Active-Control Trials: How Would a New Agent Compare with Placebo? A Method Illustrated with Clopidogrel, Aspirin, and Placebo.” *American Heart Journal* 141: 26-32.

Contains Nonbinding Recommendations

Draft – Not for Implementation

- 2284 Fleming, T.R. (1987). “Treatment Evaluation in Active Control Studies.” *Cancer Treatment*
2285 *Reports* 71, 1061-1064.
- 2286
- 2287 Fleming, T.R. (2000). “Design and Interpretation of Equivalence Trials.” *American Heart*
2288 *Journal* 139, S171-S176.
- 2289
- 2290 Follmann, D.A., Proschan, M.A. (1999). “Validity Inference in Random-Effects Meta-
2291 Analysis.” *Biometrics* 55, 732-737.
- 2292
- 2293 Freemantle, J., Cleland, J. Young, P., Mason, J., Harrison, J. (1999). “B Blockade After
2294 Myocardial Infarction: Systematic Review and Meta Regression Analysis.” *British Medical*
2295 *Journal* 318, 1730-1737.
- 2296
- 2297 Gould, A.L. (1991). “Another View of Active-Controlled Trials.” *Controlled Clinical Trials*
2298 12, 474-485.
- 2299
- 2300 Holmgren, E.B. (1999). “Establishing Equivalence by Showing That a Prespecified
2301 Percentage of the Effect of the Active Control Over Placebo is Maintained.” *Journal of*
2302 *Biopharmaceutical Statistics* 9(4), 651-659.
- 2303
- 2304 Hasselblad, V., Kong, D.F. (2001). “Statistical Methods for Comparison to Placebo in
2305 Active-Control Trials.” *Drug Information Journal* 35, 435-449.
- 2306
- 2307 Hauschke, D. (2001). “Choice of Delta: A Special Case.” *Drug Information Journal* 35,
2308 875-879.
- 2309
- 2310 Hauschke, D., Hothorn, L.A. (2007). Letter to the Editor: An Introductory Note to the
2311 CHMP Guidelines: Choice of the Non-Inferiority Margin and Data Monitoring Committees.
2312 *Statistics in Medicine* 26, 230-233.
- 2313
- 2314 Hauschke, D., Pigeot, I. (2005). “Establishing Efficacy of a New Experimental Treatment in
2315 the ‘Gold Standard’ Design (with discussions).” *Biometrical Journal* 47, 782-798.
- 2316
- 2317 Holmgren, E.B. (1999). “Establishing Equivalence by Showing that a Prespecified
2318 Percentage of the Effect of the Active Control Over Placebo is Maintained.” *Journal of*
2319 *Biopharmaceutical Statistics* 9, 651-659.
- 2320
- 2321 Hung, H.M.J., Wang, S.J., Tsong, Y., Lawrence, J., O’Neill, R.T. (2003). “Some
2322 Fundamental Issues with Non-Inferiority Testing in Active Controlled Clinical Trials.”
2323 *Statistics in Medicine* 22, 213-225.
- 2324
- 2325 Hung, H.M.J., Wang, S.J., O’Neill, R.T. (2005). “A Regulatory Perspective on Choice of
2326 Margin and Statistical Inference Issue in Non-Inferiority Trials.” *Biometrical Journal* 47,
2327 28-36.
- 2328

Contains Nonbinding Recommendations

Draft – Not for Implementation

- 2329 Hung, H.M.J., Wang, S.J., O’Neill, R.T. (2008). “Non-Inferiority Trial.” *Wiley*
2330 *Encyclopedia of Clinical Trials*. Wiley, New York.
2331
2332 Hung, H.M.J., Wang, S.J., O’Neill, R.T. (2007). “Issues with Statistical Risks for Testing
2333 Methods in Noninferiority Trial Without a Placebo Arm.” *Journal of Biopharmaceutical*
2334 *Statistics* 17, 201-213.
2335
2336 International Conference on Harmonization: *Statistical Principles for Clinical Trials* (ICH
2337 E-9), Food and Drug Administration, DHHS, 1998.
2338
2339 International Conference on Harmonization: *Choice of Control Group and Related Design*
2340 *and Conduct Issues in Clinical Trials* (ICH E-10), Food and Drug Administration, DHHS,
2341 July 2000.
2342
2343 Jones, B., Jarvis, P., Lewis, J.A., Ebbutt AF (1996). “Trials to Assess Equivalence: the
2344 Importance of Rigorous Methods.” *British Medical Journal* 313, 36-39.
2345
2346 Julious, S.A., Wang, S.J. (2008). “How Biased are Indirect Comparisons Particularly When
2347 Comparisons Are Made Over Time in Controlled Trials?” *Drug Information Journal* 42,
2348 625-633.
2349
2350 Koch, A., Röhmel, J. (2004). “Hypothesis Testing in the Gold Standard Design for Proving
2351 the Efficacy of an Experimental Treatment Relative to Placebo and a Reference.” *Journal of*
2352 *Biopharmaceutical Statistics* 14, 315-325.
2353
2354 Kaul, S., Diamond, G.A. (2006). “Good Enough: A Primer on the Analysis and
2355 Interpretation of Non-Inferiority Trials.” *Annals of Internal Medicine* 145, 62-69.
2356
2357 Lange, S., Freitag, G. (2005). “Choice of Delta: Requirements and Reality – Results of a
2358 Systematic Review.” *Biometrical Journal* 47; 12-27.
2359
2360 Laster, L.L., Johnson, M.F., Kotler, M.L. (2006). “Non-Inferiority Trials: the ‘at least as
2361 good as’ Criterion with Dichotomous Data.” *Statistics in Medicine* 25, 1115-1130.
2362
2363 Lawrence, J. (2005). “Some Remarks About the Analysis of Active Control Studies.”
2364 *Biometrical Journal* 47, 616-622.
2365
2366 Ng, T.H. (1993). “A Specification of Treatment Difference in the Design of Clinical Trials
2367 with Active Controls.” *Drug Information Journal* 27, 705-719.
2368
2369 Ng, T.H. (2001). “Choice of Delta in Equivalence Testing.” *Drug Information Journal* 35,
2370 1517-1527.
2371
2372 Ng, T.H. (2008). “Noninferiority Hypotheses and Choice of Noninferiority Margin.”
2373 *Statistics in Medicine* 27, 5392-5406.

Contains Nonbinding Recommendations

Draft – Not for Implementation

- 2374
2375 Pledger, G., Hall, D.B. (1990). “Active Control Equivalence Studies: Do They Address the
2376 Efficacy Issue?” *Statistical Issues in Drug Research and Development*, Marcel Dekker, New
2377 York, 226-238.
2378
2379 Röhmel, J. (1998). “Therapeutic Equivalence Investigations: Statistical Considerations.”
2380 *Statistics in Medicine* 17, 1703-1714.
2381
2382 Rothmann, M., Li, N., Chen, G., Chi, G.Y.H., Temple, R.T., Tsou, H.H. (2003). “Non-
2383 Inferiority Methods for Mortality Trials.” *Statistics in Medicine* 22, 239-264.
2384
2385 Rothmann, M. (2005). “Type I Error Probabilities Based on Design-Stage Strategies with
2386 Applications to Noninferiority Trials.” *J. of Biopharmaceutical Statistics* 15; 109-127.
2387
2388 Sanchez, M.M., Chen, X. (2006). “Choosing the Analysis Population in Non-Inferiority
2389 Studies: Per Protocol or Intent-to-Treat.” *Statistics in Medicine* 25, 1169-1181.
2390
2391 Sheng, D., Kim, M.Y. (2006). “The Effects of Non-Compliance on Intent-to-Treat Analysis
2392 of Equivalence Trials.” *Statistics in Medicine* 25, 1183-1190.
2393
2394 Siegel, J.P. (2000). “Equivalence and Noninferiority Trials.” *American Heart Journal* 139:
2395 S166-S170.
2396
2397 Simon, R. (1999). “Bayesian Design and Analysis of Active Control Clinical Trials.”
2398 *Biometrics* 55, 484-487.
2399
2400 Snapinn, S.M. (2004). “Alternatives for Discounting in the Analysis of Noninferiority
2401 Trials.” *Journal of Biopharmaceutical Statistics* 14, 263-273.
2402
2403 Snapinn, S.M., Jiang, Q. (2008). “Controlling the Type I Error Rate in Non-Inferiority
2404 Trials.” *Statistics in Medicine* 27, 371–381.
2405
2406 Temple, R. (1987). “Difficulties in Evaluating Positive Control Trials.” *Proceedings of the*
2407 *Biopharmaceutical Section of American Statistical Association*, 1-7.
2408
2409 Temple R. (1996). “Problems in Interpreting Active Control Equivalence Trials.”
2410 *Accountability in Research* 4: 267-275.
2411
2412 Temple, R., Ellenberg, S.S. (2000). “Placebo-Controlled Trials and Active-Control Trials in
2413 the Evaluation of New Treatments - Part 1: Ethical and Scientific Issues.” *Annals of Internal*
2414 *Medicine* 133, 455-463.
2415
2416 Wang, S.J., Hung, H.M.J., Tsong, Y. (2002). “Utility and Pitfalls of Some Statistical
2417 Methods in Active Controlled Clinical Trials.” *Controlled Clinical Trials* 23, 15-28.
2418

Contains Nonbinding Recommendations

Draft – Not for Implementation

- 2419 Wang, S.J., Hung, H.M.J. (2003). “Assessment of Treatment Efficacy in Non-Inferiority
2420 Trials.” *Controlled Clinical Trials* 24, 147-155.
2421
2422 Wang S.J., Hung H.M.J. (2003). “TACT Method for Non-Inferiority Testing in Active
2423 Controlled Trials.” *Statistics in Medicine* 22; 227-238.
2424
2425 Wang, S.J., Hung, H.M.J., Tsong, Y. (2003). “Non-Inferiority Analysis in Active Controlled
2426 Clinical Trials.” *Encyclopedia of Biopharmaceutical Statistics, 2nd Edition*. Marcel Dekker,
2427 New York.
2428
2429 Wiens, B. (2002). “Choosing an Equivalence Limit for Non-Inferiority or Equivalence
2430 Studies.” *Controlled Clinical Trials* 23, 2-14.
2431
2432 Wiens, B. (2006). “Randomization as a Basis for Inference in Noninferiority Trials.”
2433 *Pharmaceutical Statistics* 5, 265-271.