Non-Inferiority Clinical Trials to Establish Effectiveness

Guidance for Industry

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

November 2016
Clinical/Medical

Samsung v. Regeneron
IPR2023-00442
U.S. Pat. 10,130,681
Exhibit 2118

Non-Inferiority Clinical Trials to Establish Effectiveness

Guidance for Industry

Additional copies are available from:
Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10001 New Hampshire Ave., Hillandale Bldg., 4th Floor
Silver Spring, MD 20993-0002
Phone: 855-543-3784 or 301-796-3400; Fax: 301-431-6353
Email: druginfo@fda.hhs.gov
http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm
and/or
Office of Communication, Outreach and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave., Bldg. 71, Room 3128
Silver Spring, MD 20993-0002
Phone: 800-835-4709 or 240-402-8010
Email: ocod@fda.hhs.gov
http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

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

November 2016
Clinical/Medical

TABLE OF CONTENTS

I. INTRODUCTION 1

II. BACKGROUND 2

III. GENERAL CONSIDERATIONS FOR NON-INFERIORITY STUDIES 3

A. The Non-Inferiority Hypothesis 3

B. Reasons for Using a Non-Inferiority Design 7

C. The Non-Inferiority Margin 8

D. Assay Sensitivity 11

E. Statistical Inference 14

F. Regulatory Conclusions 15

G. Alternative Designs 16

H. Number of Studies Needed 17

I. Choice of Active Control 18

IV. CHOOSING THE NON-INFERIORITY MARGIN AND TESTING THE NON-INFERIORITY HYPOTHESIS 19

A. Introduction 19

B. Statistical Uncertainties and Quantification of the Active Control Effect 20

C. Fixed Margin and Synthesis Methods 27

D. Considerations for Selecting the Clinical Margin (M_2) 30

E. Estimating the Sample Size 30

F. Study Quality and Choice of Analysis Population 31

G. Testing Non-Inferiority and Superiority in a Single Trial 31

V. FREQUENTLY ASKED QUESTIONS AND GENERAL GUIDANCE 32

APPENDIX — EXAMPLES 37

REFERENCES FOR EXAMPLES 48

REFERENCES 51

Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry[1]

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

I. INTRODUCTION

This document provides guidance to sponsors and applicants submitting investigational drug applications (INDs), new drug applications (NDAs), biologics licensing applications (BLAs), or supplemental applications on the appropriate use of non-inferiority (NI) study designs to provide evidence of the effectiveness of a drug or biologic, usually because a superiority study design (drug versus placebo, dose response, or superiority to an active drug) cannot be used.[2] The guidance gives advice on when NI studies intended to demonstrate effectiveness of an investigational drug can provide interpretable results, how to choose the NI margin, and how to test the NI hypothesis.

This guidance does not provide recommendations for the use of NI study designs to evaluate the safety of a drug.

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

This guidance finalizes the draft guidance for industry, *Non-Inferiority Clinical Trials*, published in 2010. In addition, it supersedes the guidance for industry, *Antibacterial Drug Products: Use of Noninferiority Trials to Support Approval*, also published in 2010, which will be withdrawn.

[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. While most concepts discussed will be broadly applicable, certain issues related to vaccines, such as the choice of the NI margin when the study endpoint is the level of antibodies, would call for consultation from CBER.

II. BACKGROUND

FDA's regulations on adequate and well-controlled studies (21 CFR 314.126) describe four kinds of concurrently controlled trials that provide evidence of effectiveness. Three of them — placebo, no treatment, and dose-response controlled trials — are superiority trials that seek to show that a test drug is superior to the control (placebo, no treatment, or a lower dose of the test drug). The fourth kind, comparison with an active treatment (active control), can also be a superiority trial, if the intent is to show that the new drug is more effective than the control. More commonly, however, the goal of such studies is to show that the difference between the new and active control treatment is small — small enough to allow the known effectiveness of the active control, based on its performance in past studies and the assumed effectiveness of the active control in the current study, to support the conclusion that the new test drug is also effective. How to design and interpret the results of such studies so that they can support a conclusion about effectiveness of the new drug is challenging.

Active controlled trials that are not intended to show superiority of the test drug but rather to show that the new treatment is not inferior to an unacceptable extent were once called clinical equivalence trials. The intent of an NI trial, however, is not to show that the new drug is equivalent, but rather that it is not materially worse than the control. Therefore, the interest is one-sided. The new drug could be better than the control, and therefore at a minimum non-inferior, but it would not be equivalent.

The critical difference between superiority and NI trials is that a properly designed and conducted superiority trial, if successful in showing a difference, is entirely interpretable without further assumptions (other than lack of bias) — that is, the result speaks for itself and requires no extra-study information. In contrast, the NI study is dependent on knowing something that is not measured in the study, namely, that the active control had its expected effect in the NI study. When this occurs, the trial is said to have assay sensitivity, defined as the ability to have shown a difference from placebo of a specified size. A "successful" NI trial, one that shows what appears to be an acceptably small difference between treatments, may or may not have had assay sensitivity and therefore may or may not support a conclusion that the test drug was effective. Thus, if the active control had no effect at all in the NI trial (i.e., did not have any of its expected effect), then even ruling out a very small difference between control and test drug is meaningless and provides no evidence that the test drug is effective. (See Section III.D. for further discussion on assay sensitivity.) In the absence of a placebo arm, knowing whether the trial had assay sensitivity relies heavily on external (not within-study) information, giving NI studies some of the characteristics of a historically controlled trial.

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

> If the intent of the trial is to show similarity of the test and control drugs, the report of the study should assess the ability of the study to have detected a difference between treatments. Similarity of test drug and active control can mean either that both drugs were effective or that neither was effective. The analysis of the study should explain why the

drugs should be considered effective in the study, for example, by reference to results in previous placebo-controlled studies of the active control drug.

This guidance consists of four parts:

- A general discussion of regulatory, study design, scientific, and statistical issues associated with the use of NI studies to establish the effectiveness of a new drug
- A focus on some of these issues in more detail, notably the statistical approaches used to determine the NI margin and to test for non-inferiority
- Commonly asked questions about NI studies
- Four examples of successful and unsuccessful efforts to define NI margins and test for non-inferiority[3]

III. GENERAL CONSIDERATIONS FOR NON-INFERIORITY STUDIES

A. The Non-Inferiority Hypothesis

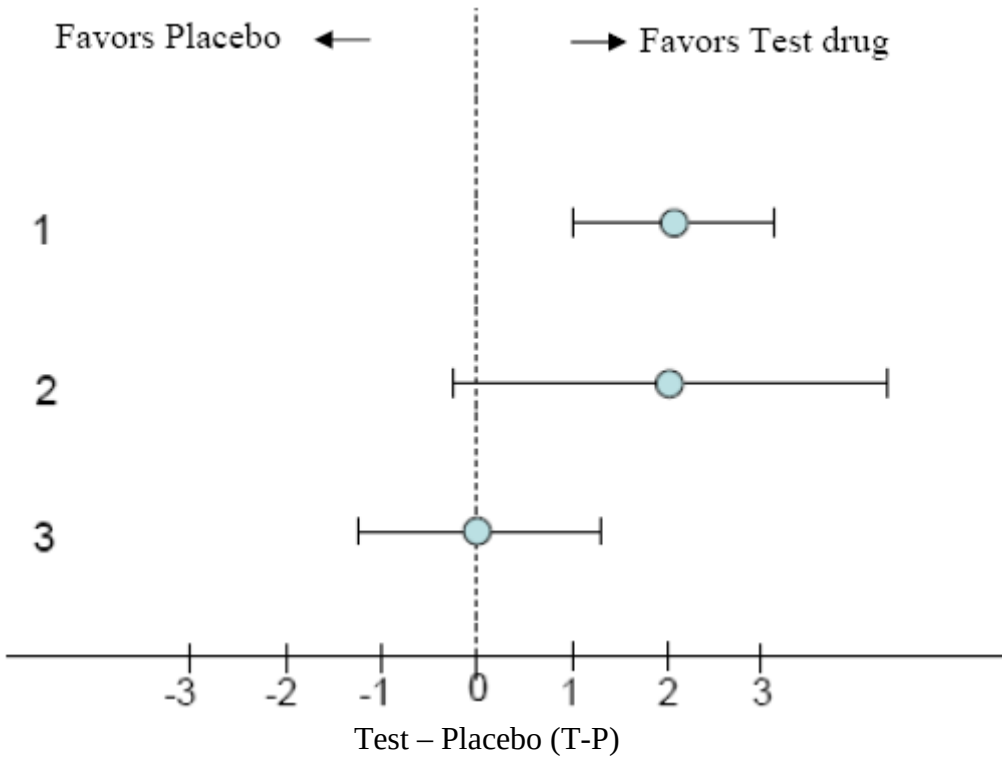
In a placebo-controlled trial, the null hypothesis (H_o) is that the beneficial response to the test drug (T) is less than or equal to the response to the placebo (P); the alternative hypothesis (H_a) is that the response to the test drug is greater than P. Thus:

$$H_o: T \leq P; \quad T - P \leq 0$$
$$H_a: T > P; \quad T - P > 0$$

In most cases, the test for a treatment effect corresponds to showing that the lower bound of the two-sided 95% confidence interval (equivalent to the lower bound of a one-sided 97.5% confidence interval) for T-P is > 0. This result shows that the effect of the test drug is greater than 0 (see Figure 1).

[3] References: In this guidance, references to methods or studies are not included in the text; rather they are included in the general references section in the Appendix and are grouped by topic. A separate references section is also provided for the examples in the Appendix.

Figure 1. Possible Results of a Placebo-Controlled Superiority Study (Point Estimate and 95% Confidence Interval (CI))



1. Point estimate of effect is 2; 95% CI lower bound is 1. Conclusion: Drug is effective and has an effect of at least 1.
2. Point estimate of effect is 2; 95% CI lower bound is <0. 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 is not shown to be effective.

Although there is no difference in the conclusions of scenarios 2 and 3, the magnitude of the treatment difference and width of the confidence interval in scenario 2 may encourage the conduct of another study (possibly larger) before deciding that the test drug has no effect.

In an NI study, the goal is to demonstrate that the test drug has an effect by showing that its effect is sufficiently close to the effect of an active control. There is no placebo arm in the study; therefore, the effect of the active control is not measured in the study but must be assumed. The goal of the study is to show that the effect of the test drug (T) is not inferior to the effect of the active control (C) by a specified amount, called the NI margin, or M.

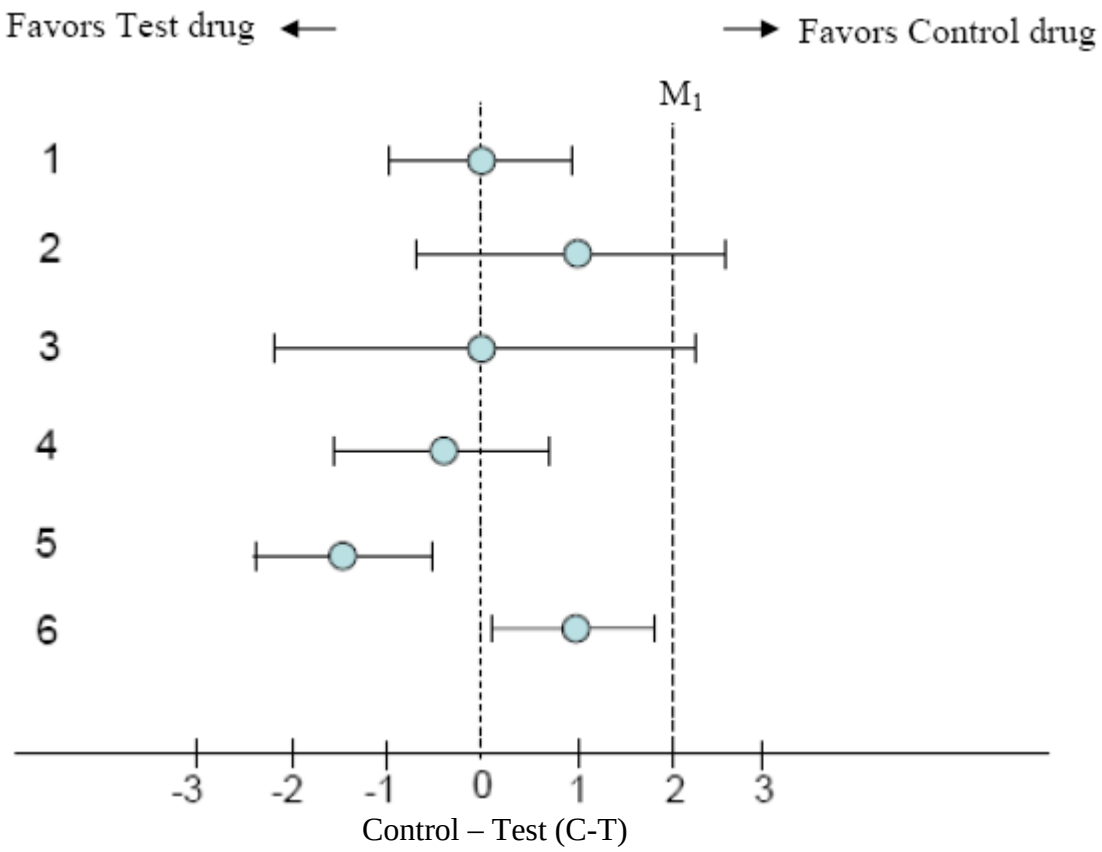
The null and alternative hypotheses correspond to a null hypothesis of inferiority and an alternative hypothesis of non-inferiority, as follows:

H_o: $C - T \geq M$ (T is inferior to the control (C) by M or more)
H_a: $C - T < M$ (T is inferior to the control (C) by less than M)

A statistical test of the above hypothesis is provided by comparing the upper bound of the two-sided confidence interval for C-T with the NI margin, M, which is specified in advance. If the upper bound is less than M, non-inferiority of T relative to C is established.

One choice for M (the largest possible value) is to set it equal to the entire known effect of the active control relative to placebo, based on past randomized controlled trials. With this choice for M, called M_1, and assuming the control drug attains this level of efficacy in the NI study, a finding of non-inferiority means that the test drug has an effect greater than 0 (see Figure 2). A more usual choice is to set M equal to some clinically relevant portion of M_1, namely, the portion of the control drug effect it is important to preserve with the test drug, based on clinical judgment.

Figure 2. Possible Results of an NI Study Showing Control Drug–Test Drug Differences (Point Estimate and 95% CI)



1. Point estimate of C-T is 0, suggesting equal effect of C and T; upper bound of the 95% CI for C-T is 1, well below M_1; NI is demonstrated.
2. Point estimate of C-T favors C; the upper bound of the 95% CI for C-T is >2, above M_1; NI is not demonstrated.
3. Point estimate of C-T is zero, which suggests an equal effect; but the upper bound of the 95% CI for C-T is >2 (i.e., above M_1), so that NI is not demonstrated.
4. Point estimate favors T; NI is demonstrated, but superiority is not demonstrated.
5. Point estimate favors T; superiority and NI are demonstrated.
6. Point estimate of C-T is 1, favoring the control. The upper bound of the 95% CI for C-T is <M_1, demonstrating NI (the entire effect of C has not been lost) but at the same time the 95% CI for C-T is above zero, indicating that T is actually inferior to C, even while meeting the NI standard.

The determination of M_1 is a critical step in designing an NI trial and is often difficult; this determination is therefore a major focus of this guidance. M_1 cannot be directly measured in the NI study, because there is no concurrent placebo group. It must be estimated based on the past performance of the active control, preferably in placebo-controlled trials, and then assumed to hold for the NI study based on a comparison of prior test conditions to the current test environment (see section III.D).

The choice of the margin M_1 has important practical consequences. The smaller the margin, the lower the upper bound of the 95% two-sided confidence interval for C-T must be, and the larger the sample size needed to establish non-inferiority. Showing that the upper bound of the 95% CI of C-T is less than M_1 demonstrates that the test drug has some effect (i.e., an effect > 0). The margin of interest, however, as noted above, is usually smaller than M_1 (to show that an adequate portion of the clinical benefit of the control is preserved), in which case it is called M_2. The basis for this expectation is described below in section III.C.4.

B. Reasons for Using a Non-Inferiority Design

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

> In cases where an available treatment is known to prevent serious harm, such as death or irreversible morbidity in the study population, it is generally inappropriate to use a placebo control. There are occasional exemptions, however, such as cases in which standard therapy has toxicity so severe that many patients have refused to receive it.
>
> In other situations, where there is no serious harm, it is generally considered ethical to ask patients to participate in a placebo-controlled trial, even if they may experience discomfort as a result, provided the setting is non-coercive and patients are fully informed about available therapies and the consequences of delaying treatment [ICH E10; pps.13-14].

Aside from this ethical reason, there may be other reasons to include an active control, possibly in conjunction with a placebo control, either to compare treatments or to assess assay sensitivity (see section III.D). Caregivers, third party payers, and some regulatory authorities have increasingly placed an emphasis on the comparative effectiveness of treatments, leading to more studies that compare two treatments. Such studies can provide information about the clinical basis for comparative effectiveness claims, which may be helpful in assessing cost effectiveness of treatments. If a placebo group is included in addition to the active comparator, it becomes possible to judge whether the study could have distinguished treatments that differed substantially, e.g., active drug versus placebo. Such comparative effectiveness studies must be distinguished from NI studies, which are the main focus of this document. The word *non-inferior* is used here in a special sense. The methods described in this document are intended to show that a new treatment that demonstrates non-inferiority is effective, not that it is as effective as the active comparator. A new treatment may meet the standard of effectiveness (that it is superior to placebo) without justifying a conclusion that it is as effective or even nearly as effective as the active comparator.

C. The Non-Inferiority Margin

As described above, the NI study seeks to show that the amount by which the test drug (T) is inferior to the active control (C), C-T, is less than some prespecified NI margin (M). M can be no larger than the presumed entire effect of the active control in the NI study, and the margin based on the entire active control effect is generally referred to as M_1. It is critical to recognize that M_1 is not measured in the NI trial (in the absence of a placebo arm), but rather is estimated based on past performance of the active control. The effect is assumed to be present in the current study based on a thorough comparison of the characteristics of the current NI study with those of prior studies and assessment of the quality of the NI study. The validity of any conclusion from the NI study depends on the choice of M_1 and its relevance to the current NI study. If, for example, the NI margin is chosen as 10, and the study does indeed rule out a difference of 10 (but not a smaller difference), seeming to demonstrate effectiveness of T, but the true effect of C in this study was actually less than 10, then a conclusion that the study demonstrated non-inferiority would have been incorrect. The choice of M_1, together with reasonable assurance that this effect occurred in the trial (i.e., the presence of assay sensitivity), is thus critical to obtaining a meaningful, correct answer in an NI study. Because assay sensitivity can never be proven in the absence of a placebo group, historical evidence of assay sensitivity is essential, but measures should be taken to make it likely that the active control will have the presumed effect in the NI study (see section III.A.5). This, together with careful selection of the NI margin, will increase the likelihood of valid and interpretable results.

Because the consequence of choosing a margin greater than the actual treatment effect of the active control in the study may be a false conclusion that a new drug is effective, an undesirable public health outcome, there is a powerful tendency to be conservative both in the choice of margin and in the statistical analysis used to rule out a degree of inferiority of the test drug to the active control of more than that margin. This is generally done by ensuring that the upper bound of the 95% two-sided confidence interval for C-T in the NI trial is smaller than M_1. Use of this interval corresponds to a one-sided test size (alpha level) of 0.025 in testing the NI hypothesis stated above. The upper bound of the confidence interval for C-T is not, however, the only measurement of interest, just as the lower bound of a 95% confidence interval for the effect of drug versus placebo in a superiority trial is not the only value of relevance. The point estimate of the treatment effect and the width of its confidence interval are also relevant. Nonetheless, the upper bound of the 95% confidence interval is typically used to judge the effectiveness of the test drug in the NI study, just as a two-sided test size (alpha level) of 0.05 is traditionally the standard used for defining success in a superiority trial. The 95% confidence interval upper bound for C-T is used, in conjunction with M_1, to provide a reasonably high level of assurance that the test drug does, in fact, have an effect greater than zero (i.e., ruling out loss of all the effect of the active control).

Although the NI margin used in a trial can be no larger than the entire assumed effect of the active control in the NI study (M_1), it is usual and generally desirable to choose a smaller value, called M_2, for the NI margin. Showing non-inferiority to M_1 would provide assurance that the test drug had an effect greater than zero, but in many cases that would not be sufficient to conclude that the test drug had a clinically acceptable effect. Recall that the main reason an NI study is conducted is that it is not ethical to include a placebo arm. The active control has a

beneficial effect, and denying that benefit to subjects with a serious illness is not ethical. For the same reason, it would not usually be acceptable for a test drug to lose most of that active control's effect. It is therefore usual in NI studies to choose a smaller margin (M_2) that reflects the largest loss of effect that would be clinically acceptable. This can be described as an absolute difference in effect (typical of antibiotic trials) or as a fraction of the risk reduction provided by the control (typical in cardiovascular outcome trials). Note that a larger value for M_2 may be justified clinically, if the test drug were shown to have some important advantage (e.g., on safety or on a secondary endpoint).

Concern has been expressed that use of M_2 represents an FDA "comparative effectiveness" standard that is not included in the Federal Food, Drug, and Cosmetic Act. In explaining the role of relative effectiveness under law in April 1995, President Clinton and Vice President Gore ("Reinventing Regulation of Drugs and Medical Devices," part of the National Performance Review) stated the following:

> In certain circumstances, however, it may be important to consider whether a new product is less effective than available alternative therapies, when less effectiveness could present a danger to the patient or to the public. For example, it is essential for public health protection that a new therapy be as effective as alternatives that are already approved for marketing when:
> 1. the disease to be treated is life-threatening or capable of causing irreversible morbidity (e.g., stroke or heart attack); or
> 2. the disease to be treated is a contagious illness that poses serious consequences to the health of others (e.g., sexually transmitted disease).

The reinvention statement was placed in the *Federal Register* of August 1, 1995 (60 FR 39180 at 39181), as an FDA position by FDA's Deputy Commissioner for Policy, William Schultz.

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

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

M_1 is estimated based on the historical experience with the active control drug. Its relevance to the current NI trial is based on:

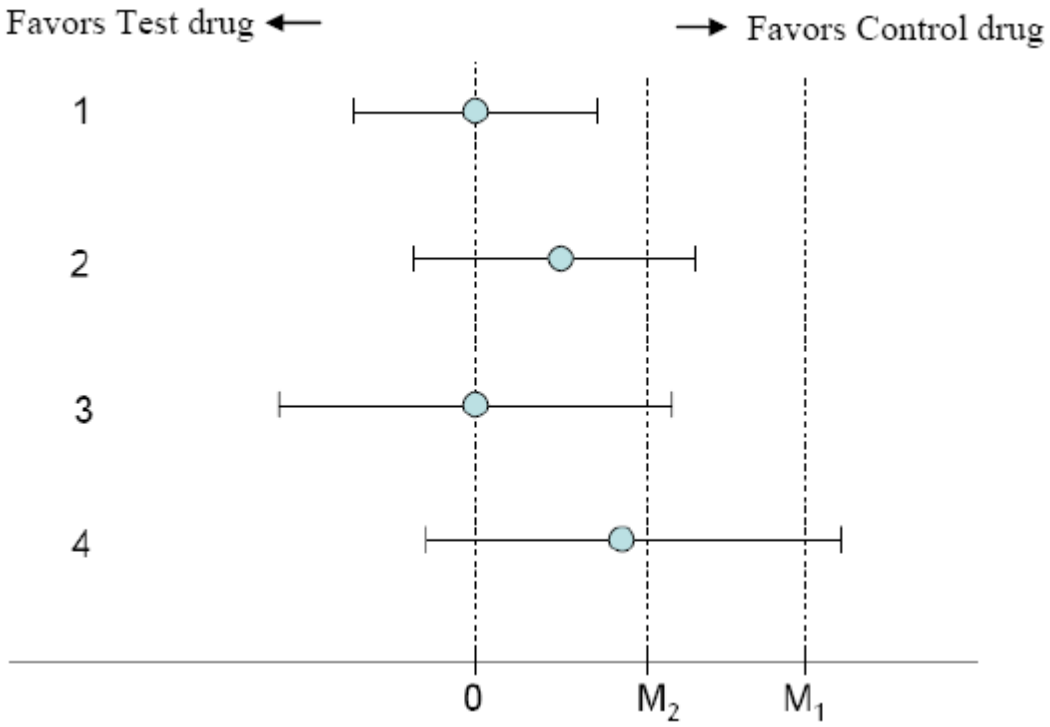
1. Assessment of the likelihood that the current effect of the active control is similar to that observed in the past studies used to estimate the active control effect (the constancy assumption)
2. Assessment of the quality of the NI trial, particularly looking for deficiencies in study design and/or conduct that could reduce a difference between the active control and the new drug

Note that in a trial seeking to show a difference (i.e., superiority), this diminution of the between-treatment difference in the second element is a "bias toward the null," but in a non-inferiority

trial, it is a "bias toward the alternative" (i.e., non-inferiority). Because of this second element, in some situations M_1, although prespecified, has to be "discounted" (i.e., a smaller value would be used), but the amount of discounting needed may not be known until the NI study has been completed.

The choice of M_2 is a matter of clinical judgment, but M_2 can never be greater than M_1, even for a situation in which the active control drug has a small effect, and clinical judgment might argue that a larger difference is not clinically important. Even if that clinical judgment were reasonable, choosing an M_2 greater than M_1 as the NI margin would not allow a conclusion that the test drug has any effect. As explained above, ruling out a difference between the active control and test drug larger than M_1 is the critical finding that supports a conclusion of effectiveness. This analysis is approached with great rigor; that is, a difference (C-T) larger than M_1 needs to be ruled out with a high degree of statistical assurance. As M_2 represents a clinical judgment, there may be a greater flexibility in interpreting a 95% upper bound for C-T that is slightly greater than M_2, as long as the upper bound is still well less than M_1 (see Figure 3).

Figure 3. Possible Results of an NI Study Showing Active Control – Test Drug Differences (Point Estimate and 95% CI)



Control – Test (C-T); (degree of inferiority of test drug)

1. C-T point estimate = 0 and upper bound of 95% CI < M_2, indicating test drug is effective and adequately rules out an unacceptable loss of the control effect (NI demonstrated).
2. Point estimate of C-T favors C; upper bound of 95% CI < M_1 but > M_2, indicating test drug effect > 0 but an unacceptable loss of the control effect has not been ruled out.
3. Point estimate of C-T is zero and upper bound of 95% CI < M_1 but it is slightly greater than M_2. Loss of the pre-specified M_2 has thus not been ruled out, but whether the study has shown adequate preservation of the control effect would be a matter of clinical judgment.
4. C-T point estimate favors C and upper bound of 95% CI > M_1, indicating the study does not provide evidence of effectiveness for test drug.

D. Assay Sensitivity

Assay sensitivity is an essential property of an NI clinical trial. Assay sensitivity is the ability of the trial to have detected a difference between treatments of a specified size. Stated in another way, assay sensitivity means that had the study included a placebo, a control drug-placebo difference of at least M_1 would have been present.

As explained above, the choice of M_1, and the conclusion that a trial has assay sensitivity (i.e., the active control would have had an effect of at least M_1), is based on three considerations:

1. Historical evidence of sensitivity to drug effects
2. The similarity of the new NI trial to the historical trials (the constancy assumption)
3. The quality of the new trial (ruling out defects that would tend to minimize differences between treatments)

1. Historical Evidence of Sensitivity to Drug Effects (HESDE)

HESDE means that prior studies, which were appropriately designed and conducted trials in the past and 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). Consistent findings in past studies allow for a reliable estimate of the drug's effect compared to placebo. The estimate of the size of the effect must take the variability of past results into account; one should not presume that the largest effect seen in any trial, or even the point estimate of a meta-analysis of all relevant trials, is certain to be repeated. Analysis of historical data will be discussed further in section IV.

HESDE cannot be determined for many symptomatic treatments (e.g., treatments for depression, anxiety, insomnia, angina, symptomatic heart failure, symptoms of irritable bowel disease, and pain), as even well-designed and conducted studies for such indications often fail to distinguish an effective drug from placebo. In those cases, it cannot be assumed 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 indications may therefore be non-informative. These issues can also negatively affect the assessment of effectiveness in outcome studies. For example, in the case of aspirin, the largest placebo-controlled trial (AMIS, the Aspirin Myocardial Infarction Study; see Example 2) did not show an effect of aspirin even though other trials all favored aspirin. Similarly, of more than 30 post-infarction beta-blocker trials, only a small number showed significantly improved survival or other cardiovascular benefit.

2. Similarity of the Current NI Trial to the Historical Studies and Its Relationship to the "Constancy Assumption"

The conclusion that HESDE can be used to estimate the effect of the active control, which will then serve as the basis for choosing M_1 for the new NI study, can be reached only when it is appropriate to conclude that the NI study is sufficiently similar to the past studies with respect to all important study design and conduct features that might influence the active control effect. This conclusion is referred to as the "constancy assumption." The design features of interest include:

- The characteristics of the patient population
- Important concomitant treatments
- Definitions and ascertainment of study endpoints
- Dose of active control

- Entry criteria
- Analytic approaches

For example, the effect of an angiotensin-converting enzyme (ACE) inhibitor on heart failure mortality has repeatedly been shown in studies where the drugs were added to diuretics and (frequently) digoxin, establishing HESDE, but evolution in treatment since those studies were conducted raises questions about our understanding of the present-day effect of ACE inhibitors. Since the time of those studies, other medications (beta blockers, spironolactone) have come into standard use. We do not know whether the past effect would still be present when ACE inhibitors are added to a regimen including drugs from these two classes. Similarly, the effect of a thrombolytic on cardiovascular mortality could depend on how soon after symptoms the drug was given, concomitant use of anticoagulants and platelet inhibitors, and use of lipid-lowering drugs. To provide a sound basis for choosing M_1, the historical studies and the new NI study should be as nearly identical as possible in all important respects.

Providing reasonable assurance that endpoints in the historical trial will be similar to, and will have been evaluated similarly to, endpoints in the new trial is easier when the endpoints are standardized and objective. The effect of the active control could be on a single endpoint (e.g., mortality) or on a composite (e.g., death, heart attack, and stroke), but, again, it is critical that measurement and assessment of these be reasonably consistent over time. The endpoint used in the NI study need not necessarily be the one used in the original trials of the active control if data from the historical studies are available to estimate the size of the effect of the active control on the new endpoint used in the NI study. For example, even if the historical studies used a mortality endpoint, the studies could be used, if high quality data could be obtained, to calculate the magnitude of an effect for an endpoint of death plus hospitalization, as long as it was possible to be confident that the circumstances leading to the hospitalization were similar in the historical studies and the NI study. Note, however, that it would not be acceptable to search through a range of endpoints to find the largest historical effect, as this could represent an overestimate of the effect to be expected in the NI study.

In general, where there has been substantial evolution over time in disease definition and treatment, or where the methodology used in the historical trials has become outdated, the assumption of constancy may not be supported, and, therefore, the use of an NI design may not be justified.

Although an NI study can be designed to be similar in most aspects to the historical studies, it may not be possible to assess that similarity fully until the NI study is completed and various characteristics of the study population and response are evaluated. When there is known heterogeneity of the active control treatment effect related to patient characteristics (e.g., age, gender, severity) and that heterogeneity can be quantified, it may be necessary to adjust the estimate of the size of the active control effect in the NI study if the mix of patient characteristics in the historical studies and the NI study differ substantially.

The property of constancy of the treatment effect may depend on which metric is chosen to represent the treatment effect. This issue is discussed in more depth in section IV.B.2.c. Experience suggests that when background rates of outcome events differ among study

populations, metrics like hazard ratio or relative risk may be more stable than metrics like absolute risk difference, which are more sensitive to changes in event rates in the population.

3. Good Study Quality

Conducting any poor quality studies should always be avoided, but with NI studies, sloppiness in study design/conduct is particularly problematic, because it introduces bias towards the alternative hypothesis of non-inferiority (see ICH E10; pp. 11-12 and section IV.F for further discussion). Deficiencies such as imprecise or poorly implemented entry criteria, poor compliance, use of concomitant treatments whose effects may overlap with the drugs under study, inadequate measurement techniques, errors in delivering assigned treatments, high attrition, or poor follow-up may reduce the difference C-T observed in the study, potentially leading to a false conclusion of non-inferiority. It should also be appreciated that intent-to-treat approaches, which preserve the principle that all patients are analyzed according to the treatment to which they have been randomized even if they do not receive it, although conservative in superiority trials, are not necessarily conservative in an NI study and can lead to an incorrect finding of non-inferiority.

In a superiority trial, sloppiness can lead to study failure. In contrast, poor quality in an NI trial can sometimes lead to an apparent finding of non-inferiority that is incorrect. Therefore, particular attention to quality is critical when planning and conducting an NI study. Adjustment for poor quality after the fact is usually not possible.

E. Statistical Inference

The various approaches to calculating the NI margin and analyzing an NI study will be discussed in detail in section IV. A commonly used fixed margin method is generally referred to as the *95%-95% method*. The first 95% refers to the confidence interval of the estimated effect of the control based on the historical studies demonstrating the effect, and the second 95% refers to the confidence interval used to test the null hypothesis in the NI study. Note that the first 95% lower confidence bound discussed here is a bound for the *average* effect of the active comparator in historical studies, or the true effect in the NI study if the constancy assumption holds. It is sometimes supposed that this is a lower bound for the actual effect of the comparator in the NI study, and the draft guidance (issued in 2010) suggests this, but this is not correct. The 95% lower bound does indeed define a lower bound for the true control effect, but the actual effect in the NI study would still be liable to sampling variation. To bound the actual effect, a different type of interval would be required; these intervals are known as *prediction intervals* and are much wider than the corresponding confidence intervals. Confidence intervals nevertheless suffice for the present purpose, because the confidence interval from the historical studies bounds the true effect of the comparator relative to placebo, and the confidence interval from the NI study bounds the true effect of the test drug relative to the comparator. Relying always on the constancy assumption, these can be combined to infer a positive effect of the test drug relative to placebo.

The 95%-95% fixed margin approach (using in this case a risk difference rather than a risk ratio for the margin) can be illustrated by FDA's evaluation of a new thrombolytic product, reteplase,

for treatment of acute myocardial infarction (AMI). To calculate the NI margin, results of all available placebo-controlled trials of streptokinase, the active comparator (control) for the NI study, were pooled by means of an appropriate meta-analytic method. The pooled results provided a point estimate for the effect on survival of an absolute 2.6% difference in mortality rates, with a 95% lower bound of 2.1% (i.e., M_1). A clinical decision was made that any new thrombolytic should rule out a loss of more than half of the benefit of streptokinase to be regarded as an acceptable alternative. The NI study would therefore have had to rule out an absolute 1.05% increase in mortality rate (i.e., M_2) in the reteplase-treated patients compared to those treated with streptokinase. The NI analysis for this study was to show that the 95% confidence interval (one-sided for this particular case) of the difference in mortality rates excluded an increase of 1.05%. The INJECT study accomplished this, and the product was approved for marketing.

An alternative to the fixed margin approach is known as the synthesis approach because it combines or synthesizes the data from the historical trials and the current NI trial, reflecting the variability in both data sources. Although the 95%-95% method was developed in a different way, it has been noted that it can be viewed as mathematically equivalent to the synthesis method but with the standard error of the effect of the test drug relative to placebo, estimated by

$$SE_H + SE_N ,$$

where SE_H *and* SE_N are the standard errors from the historical studies and the NI study, respectively, rather than by

$$\sqrt{SE_H^2 + SE_N^2} ,$$

which is the standard error for the synthesis method. The first formula always gives a larger standard error than the second. Thus, compared to the synthesis method, the 95%-95% method is conservative. Both methods rely crucially on the assumption of constant effect of the control (on average) between the historical studies and the NI study. The use of the 95%-95% method might be seen as an allowance for possible deviation from this assumption, a desirable feature for evaluating whether loss of M_1 (the whole effect of the control) has been ruled out. Use of the synthesis method, along with a careful justification of the constancy assumption and, when appropriate, an explicit allowance for deviation from it, would also be acceptable and may be recommended for determining whether a loss of effect greater than M_2 has been ruled out (see Section IV.C and Example 1(B) for more detail).

F. Regulatory Conclusions

A successful NI study shows rigorously that the test drug has an effect greater than zero if it excludes an NI margin of M_1, as long as M_1 is well chosen and represents an effect that the control drug actually would have had (versus a placebo, had there been a placebo group). The NI study can also be used to show that the test drug had an effect greater than some fraction of the control drug effect, depending on the M_2 that is used. Note, however, that although a successful NI study supports effectiveness of the test drug, it will only rarely support a conclusion that the drug is "equivalent" or "similar" to the active control, a concept that has not

been well defined for these situations. It should be appreciated that in addition to the rigorous demonstration of effectiveness based on significance testing, the trial provides additional information, just as a placebo-controlled trial supporting the effectiveness of a drug does. The point estimate of the drug effect and its confidence interval provide information about how large the difference between the test and control drug is likely to be.

As noted before, a successful NI study will usually not be sufficient to support a conclusion that the test drug is equivalent or similar to the active control. The methods discussed in this document, especially with regard to choosing the margin, are not intended to demonstrate equivalence or similarity but only to show that the test drug is effective. However, if the lower bound of the confidence interval for the effect of the test drug relative to the active control were only slightly negative, a judgment that similarity had been shown would be possible. If studies are intended to support equivalence, the margin against which the confidence interval is to be judged should be justified in advance, and the margin will usually be smaller than M_2.

G. Alternative Designs

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

1. Add-on Study

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

2. Identifying a Population Not Known to Benefit From Available Therapy in Which a Placebo-Controlled Trial Is Acceptable

In many outcome study settings, effectiveness is established for some clinical settings (e.g., severe disease) but not others. Therefore, it may be possible to study less severely ill patients in placebo-controlled trials. The demonstration that simvastatin was effective in hypercholesterolemic post-infarction patients (4S), for example, did not preclude studies of statins in hypercholesterolemic non-infarction patients (WOSCOPS) or in patients with lesser degrees of hypercholesterolemia (TEXCAPS). This approach is appropriate as long as there is uncertainty as to whether the treatment is of value in the new study population. It may also be possible to study patients intolerant to the known effective therapy. For example, it was possible to study angiotensin receptor blockers in a placebo-controlled trial in heart failure patients intolerant of ACE inhibitors, but it would not have been possible to deny a more general

population of heart failure patients an ACE inhibitor, as it had already been established that ACE inhibitors improved survival in a general heart failure population.

3. Early Escape, Rescue Treatment, Randomized Withdrawal

In symptomatic conditions, there may be reluctance to leave people on placebo for prolonged periods when effective therapy exists. It is possible to incorporate early escape/rescue provisions for patients who do not respond by a particular time, or to use a design that terminates patients on first recurrence of a symptom such as unstable angina, grand mal seizure, or paroxysmal supra-ventricular tachycardia. To evaluate the persistence of effects over time, where conducting a long-term placebo-controlled trial would be difficult, a randomized withdrawal study can be used. In these studies, patients successfully treated with a drug are randomly assigned to placebo or continued drug treatment. As soon as symptoms return, the patient is considered to have had an endpoint.

H. Number of Studies Needed

Ordinarily, with exceptions allowed by the FDA Modernization Act of 1997 (the Modernization Act), FDA expects that there will be more than one adequate and well-controlled study supporting effectiveness. The Modernization Act allows one study plus confirmatory evidence to serve as substantial evidence in some cases, and FDA has issued a guidance that discusses when a single study might be sufficient.[2]

Where there is uncertainty about the magnitude of the historical treatment effect (and thus M_1) because of variability or reliance on a single historical study, more than one NI study is usually needed to support effectiveness.

Where the studies are of relatively modest size, there is usually no impediment to conducting more than one NI trial if that appears necessary. Conducting two trials that are very large (to have adequate statistical power), however, may be infeasible, and it is worth considering what might make a single trial persuasive. Generally, two considerations might do so: (1) availability of other relevant information and (2) a statistically persuasive result.

1. Other Relevant Information

In NI trials, the test drug is generally pharmacologically similar to the active control (i.e., if they were not pharmacologically similar, an add-on study would usually have been more persuasive and more practical). In these cases, the expectation of similar performance (but still requiring confirmation in a trial) might make it possible to accept a single trial and perhaps could also

[2] See the guidance for industry *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products (*Providing Clinical Evidence of Effectiveness guidance*)*, available on the FDA Drug Web page at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm. We update guidances periodically. To make sure you have the most recent version of a guidance, check the FDA Drug Web page.

allow less conservative choices in choosing the NI margin. A similar conclusion might be reached when other types of data are available, for example:

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

2. Statistically Persuasive Result

A conclusion that an NI trial can be considered statistically persuasive may be based on the internal consistency of the NI finding or on the margin that is ruled out with a two-sided 95% confidence interval. It is important to recognize that there are two margins of interest, M_1 and M_2. In an NI study, the clinically determined margin M_2 is smaller, often considerably smaller, than M_1, which is used to determine whether the test drug has any effect. For example, M_2 might be chosen to be 40% of M_1. By meeting this M_2 criterion, ruling out a loss of 40% of the effect of the control, a single NI study provides reasonable assurance that the test drug preserves a clinically sufficient fraction (at least 60%) of the effect of the control treatment. At the same time, it provides strong assurance that the test drug has an effect greater than zero. In such cases, a single trial would usually be a sufficient basis for approval. Where the preservation of effect is particularly critical, of course, demonstrating that loss of M_2 has been ruled out in more than one study might be considered necessary.

In some cases, a study planned as an NI study may show superiority to the active control. Recommendations in International Conference on Harmonisation guidance E9: *Statistical Principles for Clinical Trials* (ICH E9) and FDA policy have been that this superiority finding arising in an NI study can be interpreted without adjustment for multiplicity. Showing superiority to an active control is a very persuasive demonstration of the effectiveness of the test drug, because demonstrating superiority to an active drug is much more difficult than demonstrating superiority to placebo. Similarly, a finding of less than superiority, but with a 95% confidence interval upper bound for C-T considerably smaller than M_2, is also statistically persuasive for demonstrating effectiveness.

I. Choice of Active Control

The active control must be a drug whose effect is well-defined. The most obvious choice is a single drug for which historical placebo-controlled trials are available to define the active control effect assumed for the NI trial. When there are several pharmacologically similar drugs, it may be possible to use a combined analysis to choose an NI margin, and in that case, any one of the related drugs may be considered as the active control in the NI trial (see section IV.B.2.b).

IV. CHOOSING THE NON-INFERIORITY MARGIN AND TESTING THE NON-INFERIORITY HYPOTHESIS

A. Introduction

In this section, various sources of uncertainty that come into play for NI trials are discussed, with particular emphasis on those affecting the choice of margin. We also discuss in more detail the choice of statistical test for the NI hypothesis. As described briefly in section III, there are two different approaches to analysis of the NI study, one called the *fixed margin method* and the other called the *synthesis method*. Both use the same data from the historical studies and NI study, but in different ways.

● With the fixed margin method, the margin M_1 is based upon estimates of the effect of the active comparator in previously conducted studies. The NI margin to be ruled out in the NI study is then prespecified, and it is usually chosen as a quantity smaller than M_1 (i.e., M_2) to ensure that a reasonable fraction of the effect of the control is preserved. The NI study is successful if the results rule out with a sufficient level of confidence (e.g., 97.5%) inferiority of the test drug to the control by an amount equal to the NI margin or more. It is referred to as a fixed margin method because the past studies comparing the control drug with placebo are used to derive a single fixed value for M_1. The value typically chosen is the lower bound of the 95% confidence interval about the treatment effect of a single placebo-controlled trial or meta-analysis of such trials and represents a conservative estimate of the effect the active control drug is expected to have in the NI study. If M_1 is ruled out by the 95% confidence interval upper bound for C-T, then the conclusion is made that the test drug has an effect. If M_2 is ruled out, then the effect of the test drug has been shown to preserve a clinically important fraction of the effect of the control drug.

● The synthesis method, derived from the same data, combines (or synthesizes) the estimate of treatment effect relative to the control from the NI trial with the estimate of the control effect from a meta-analysis of historical trials. This method treats both sources of data as if they came from the same randomized trial, to project what the placebo effect would have been had the placebo been present in the NI trial. The process makes use of the variability from both the NI trial and the historical trials and yields one confidence interval for testing the NI hypothesis that the treatment rules out loss of a prespecified fixed fraction of the control effect without actually specifying that control effect or a specific fixed NI margin based on the control effect. In other words, the fraction M_2/M_1 of the control effect that is to be preserved in evaluating NI is specified in advance, but neither quantity (M_1 or M_2) is specified.

The multiple steps involved and assumptions made in evaluating the NI hypothesis using either the fixed-margin or synthesis approach are all potential sources of uncertainty that may be introduced into the results and conclusions of an NI study. In section IV.B, we attempt to identify these sources and suggest approaches to accounting for uncertainty to reduce the possibility of drawing false conclusions from an NI study. Section IV.C provides further discussion of the statistical analysis methods for evaluating non-inferiority.

B. Statistical Uncertainties and Quantification of the Active Control Effect (M_1)

1. What Are the Sources of Uncertainty in an NI Study?

The interpretation of an NI study depends on three linked critical conclusions:

1. That there is reliable information about the effect the active control drug had in past studies
2. That there is reason to believe the effect the active control drug has in the current NI study is similar to the effect observed in past studies
3. That the NI study provides reliable information about the effect of the test drug relative to the comparator

All three sources of information are subject to uncertainty. For the first and third, the uncertainty is largely of a statistical nature, measured by standard errors (for the synthesis method) and confidence intervals (for the fixed margin method). The second conclusion is subject to scientific uncertainty that is largely unquantifiable.

The first source of uncertainty is the estimation of the effect of the active control in past studies. Particular problems arise when there is only a single historical study of the active control versus placebo because there is then no information about study-to-study variability; when there are multiple studies but substantial inconsistency in the estimates of the size of the effect among them; and when data from several pharmacologically related drugs are used to develop the estimate for the effect of the active control. All three of these potential problems need to be considered when choosing M_1.

The second source of uncertainty is not statistically based but rather arises from the concern that the magnitude of effect estimated from past studies will be larger than the effect of the active control in the current NI study. Assuming that the effect will be unchanged is often referred to as the "constancy assumption." If the assumption is incorrect and the magnitude of the effect in the current NI study is smaller than the estimated effect from historical studies, M_1 will have been incorrectly chosen (too large) and an apparently successful study showing NI could have given an erroneous result. Lack of constancy can occur for many reasons, including advances in adjunctive medical care, differences in the patient populations, or changes in the assessment of the endpoints under study. As discussed in section IV.B.2.c, there is some experience to support the view that in cardiovascular outcome studies, the absolute size of the treatment effect is more likely to be variable and sensitive to the background rates in the control group than is the risk reduction. The risk reduction may thus be a more constant measure of control drug effect than the absolute effect. How to adjust the NI margin for concerns about constancy is inevitably a matter of judgment.

The third source of uncertainty involves the risk of making a false conclusion based on the results of the test of the NI hypothesis in the NI study (i.e., concluding that $C\text{-}T < M_1$ when it is not). This uncertainty is referred to as the *Type I error*, or the false positive conclusion risk, and is similar to the concern in a placebo-controlled superiority trial that one might mistakenly conclude that a drug is more effective than placebo. It is, in other words, present in any

hypothesis-testing situation. In the NI case, the statistical test is intended to ensure that the difference between control and test drug (C-T, the degree of superiority of the control over the test drug) is smaller than the NI margin, meaning that some of the effect of the control is preserved (if C-T < M_1) or that a sufficient amount is preserved (if C-T < M_2). Typically, the one-sided Type I error is set at 0.025 by asking that the upper bound of the 95% confidence interval for C-T be less than the NI margin; this is roughly similar to the usual statistical test of superiority in a placebo-controlled trial. If only one NI study is going to be conducted, the probability of a Type I error can be made smaller by requiring that the upper bound of a confidence interval greater than 95% (e.g., 99%) be calculated and be less than the margin. This approach is similar to what is a commonly done for a single placebo-controlled trial (e.g., testing at an alpha of 0.01 - 0.001 instead of 0.05). As noted earlier, however, there may be prior information that eases this concern, and a single study at the usual Type I error boundary (0.025) may be considered sufficient if, for example, the drug and active control are pharmacologically similar.

In the following subsections, we elaborate on the impact of the first two sources of uncertainty as well as the choice of margin to use in hypothesis testing.

2. Quantification of the Active Control Effect

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

The magnitude of the treatment effect of the active control may be determined in several ways, depending upon the amount of data and the number of separate studies of similar design available to support this determination. Having many independent historical studies is generally more informative than having only one or two, because the estimate of the active control effect can be more precise and less subject to uncertainty, and because it becomes possible to judge the constancy of the effect for at least the time period spanned by the studies.

a. Determining HESDE from a single study

The most common situation in which an NI design is used to demonstrate effectiveness involves outcome studies where the active control drug has been approved for use to reduce the risk of major events (e.g., death, stroke, heart attack, continued infection, or tumor progression). It is not unusual for such approval to have been based on a single study in a specific setting, although there may be other pertinent data in related conditions or in different populations, or with pharmacologically similar drugs. Generally, basing an NI margin on a single randomized placebo-controlled superiority study would need to take into account the variability of the data in that study. The estimate of the treatment effect is usually represented by some metric such as the difference between the event rate in the active treatment group and the placebo control group,

which can be a difference in event rates or a risk ratio. The treatment effect has an uncertainty that is usually represented by the confidence interval. The lower bound of the 95% confidence interval provides a conservative estimate of the effect the active control had in the historical study and can be assumed to have in a future NI study, recognizing the possibility that changes in practice could reduce the historical effect somewhat. It is critical that a conservative estimate of the effect of the active control be used as the basis of the NI margin, because the logic of the NI study depends on the assumption that the active control has an effect at least equal to M_1 in the NI study.

If the p-value of the estimated treatment effect is much smaller than 0.05, for instance in the range of 0.01 or 0.001 or even smaller, the lower bound of the 95% confidence interval will generally be well above zero (in absolute value) or well below 1.0 (for hazard ratio and other risk estimates). In this case, we are more certain that the treatment effect is real and that the effect of the control in the NI study will be of reasonable size.

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

b. Determining HESDE from multiple trials

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

Multiple studies of the active control relative to a placebo, ineffective treatment, or no treatment provide an opportunity to obtain not only an estimate of the average effect of the active control and its uncertainty, but also a measure of the study-to-study variability of the effect. Study-to-study variability in the active control effect is a critical consideration because one of the basic assumptions in NI studies is the consistency of the effect between the historical studies and the current NI study.

Several cases illustrate the use of multiple studies to estimate the active control effect and problems that can arise.

- The ideal case is one where (1) there are a number of studies, each of sufficient size to demonstrate the effect of the active control and (2) the effects derived from these studies are reasonably consistent, so that an average effect derived from a meta-analysis provides a reasonable estimate of the control effect and supports a credible choice of M_1.

- If there are many small studies, some of which have not demonstrated an effect of the active control, it may still be possible to estimate the active control effect combining the studies using appropriate meta-analysis methods, but care should be taken if there is evidence of a large amount of study-to-study heterogeneity in the treatment effects.

- If there are multiple large studies, it would be troublesome if one or more of the studies showed little or no effect. Unless there is a convincing explanation as to why one of the studies did not demonstrate an effect, a failed study may argue against use of an NI design.

- There are sometimes several trials of different drugs in the same pharmacologic class. Pooling them with appropriate meta-analytic methods may allow calculation of a 95% confidence interval that is narrower than that from any single study. The presumption that the pharmacologically similar drugs would have similar effects may be reasonable, but care should be exercised in extending this assumption too far. If the effects of these different drugs vary considerably in the trials, it may be reasonable to use the combined data to estimate the average effect with a meta-analysis and then select the drug with the largest effect (point estimate) as the active control in the NI study.

When multiple studies of the active control are available, it is important to consider all studies and all patients when estimating the average effect. Dropping a study that does not show an effect, unless there is a very good reason, can overestimate the control drug effect and give a falsely high M_1. As noted above, the existence of properly designed and sized studies that show no treatment effect of the active control may preclude conducting NI studies with that control unless there are valid reasons to explain the disparate results.

When combining data from multiple studies to estimate the active control effect, some variation of effect is expected. Random-effects meta-analysis can be used that allow for both between-study and within-study heterogeneity. Both frequentist and Bayesian methods are available that differ in the assumptions made about the distribution of the random effects. These methods should be used to explore whether the variation between studies is more than would be expected by chance. If so, the between-study variation should be taken into account and will result in a more conservative margin for the NI study.

Examples 1 and 2 in the Appendix illustrate in more detail how multiple historical placebo-controlled trials of the active control are evaluated in designing an NI trial.

c. Metrics

As previously discussed, the assumption of constancy of the effect of the active comparator between the historical studies and the NI study is crucial to a conclusion of efficacy based on non-inferiority. There are, however, technical difficulties in defining what is meant by a constant effect across studies in different populations and at different times. The mathematical way in which the treatment effect is formulated will affect the meaning and plausibility of the constancy assumption.

Consider, for example, studies of fairly infrequent, binary outcomes, typical of most cardiovascular studies, a situation in which NI trials have been successfully employed. Suppose the chance of an event in the placebo group of a historical study was p_1, and in the active treatment group p_2. Then the treatment effect in the historical study can be expressed in several ways:

1. The ratio p_2/p_1, called the relative risk or risk ratio
2. The relative risk reduction $1 - p_2/p_1$
3. The odds ratio $[p_2/(1 - p_2)]/[p_1/(1 - p_1)]$
4. The risk difference $p_1 - p_2$
5. The number needed to treat (NNT), $1/(p_1 - p_2)$

If p_1 and p_2 vary across studies, they might nevertheless vary in such a way that p_2/p_1 is constant. For example, if p_1 and p_2 are 0.10 and 0.05 in one study, and 0.06 and 0.03 in another study, the ratio p_2/p_1 is 0.5 in both studies. The relative risk reduction will then also be the same in both studies, and the odds ratio will be approximately the same. The risk difference, however, is 0.05 in one study and 0.03 in the other, and the numbers needed to treat are 20 and 33.

In contrast, if two studies have different p_1 and p_2 but the same risk difference, they will have the same NNT but cannot have the same relative risk, relative risk reduction, or odds ratio. Thus, the very definition of *constancy* depends on how the effect is formulated mathematically. Because the interpretation of a NI study relies crucially on the assumption of constancy, careful consideration must be given to the question of what, if anything, is likely to be constant across historical studies and between the historical studies and the NI study. As noted, based on both experience and expectation, the constancy assumption for outcome studies has generally been based on expected constancy of relative and not absolute effects.

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

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

This approach requires that the control drug have an effect of at least 10% relative to placebo (had there been one) in the NI study. If the population in the NI study did not have such a benefit (e.g., if the patients all had illnesses not susceptible to the test drug such that the benefit was less than 10%), then even if the 5% difference were ruled out, that would not demonstrate the desired effectiveness (although it would seem to). Note that in this case, if the true effect of the control in the study were 8%, then ruling out a 5% difference would in fact show some effect of the test drug, just not the desired 50% of control effect.

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

Metrics like the risk ratio may be less affected by variability in the event rates in a placebo group that would occur in a future study. For example, a risk ratio for the event of interest of 3/4 =0.75 can be derived from very different absolute failure rates from different studies, as shown in the table below. While the risk ratio is similar in all four hypothetical studies, the absolute difference in failure rates ranges from 5% to 20%. Suppose that the NI margin were based on historical studies showing control drug effects like those in the fourth study. The NI margin (M_1) would then be chosen as 20%. Now suppose that under more modern circumstances the NI study had a control rate more like Study 1 and the size of the treatment effect compared to placebo is far less than 20%. An NI margin (M_1) of 20% would then be far greater than the drug effect in the NI study, and ruling out a difference of 20% would not demonstrate effectiveness at all. Thus, if the NI margin were chosen as ruling out an inferiority of 33% (or a relative risk of 1.33, i.e., 1 ÷ 0.75), and the control rate were 15%, the difference (M_1) between test and control would need to be less than 5% (15% x 1.33 = 20%, or 5% > the 15% rate in the active control group).

Study Number	Risk Ratio (C/P)	Control rate	Placebo rate
Study 1	¾	15%	20%
Study 2	¾	30%	40%
Study 3	¾	45%	60%
Study 4	¾	60%	80%

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

Revisiting the example from section III.E, it was clear in the placebo-controlled trials of streptokinase that the mortality rate in the placebo group was decreasing, so that even if the mortality reduction were constant as a relative reduction, the absolute reduction would decrease and the expected size of the treatment effect (M_1) would no longer be an absolute 2.1% reduction. The relative risk of mortality (ratio of mortality rates) was therefore used for subsequent thrombolytic product development programs because thrombolytic efficacy expressed as the relative mortality rates appeared to be more stable over time as improved standards of care were provided to all patients in a study.

For the development of tenecteplase in the treatment of AMI, evaluation of the relevant historical studies led to the conclusion that the comparator (90 minute infusion of alteplase) would reduce the relative risk of mortality compared to placebo by at least 0.22 (to about 0.78, upper limit of the relative risk confidence interval). The NI margin was chosen, again based upon clinical judgment, so the new drug would retain at least half of the efficacy of the comparator (M_1 = 0.22), i.e., at least 0.11 (M_2). The relative mortality risk of tenecteplase compared to placebo should be not greater than 0.89. This provided a value for the limit of the relative risk comparison of tenecteplase to alteplase of 1.14 ([tenecteplase/placebo]/[alteplase/placebo] = 0.89/0.78), which had to be excluded by the 95% confidence interval (one-sided in this case) of the ratio of mortality rates in the NI study. The ASSENT II study results provided a 95% one-sided upper limit of relative risk of 1.104, which led to the marketing approval of tenecteplase.

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

d. Discounting

One strategy employed in choosing the margin M_1 for the NI study design is "discounting" or reducing the magnitude of the active control's effect that is estimated from the HESDE analysis. Such discounting is done to account for the uncertainties in the assumptions that need to be made in estimating, based on past performance, the effect of the active control in the NI study. This concept of discounting focuses on M_1 determination and is distinct from a clinical judgment that the effect that can be lost on clinical grounds should be some fraction of M_1 (i.e., M_2). As discussed above, there are uncertainties associated with translating the historical effect of the active control (HESDE) to the new situation of the active control NI trial, and it is tempting to deal with that uncertainty in the constancy assumption by discounting the effect (e.g., "take half"). Rather than applying "automatic" discounting in choosing M_1, concerns about the active control effect should, to the extent possible:

- Be specific
- Make use of available data (e.g., magnitude of possible differences in effect in different patient populations, consistency of past studies, and consistency within studies across population subsets)
- Take into account factors that reduce the need for a conservative estimate, such as the pharmacologic similarity of the test and control drugs and pharmacodynamic effects of the new drug

A closely related issue is the possible need for adjustment of M_1 to reflect any differences that were observed between the population in the NI study and the historical study. For example, a finding of a smaller effect in women than in men in the historical studies would need to be considered in assessing the validity of M_1 if the NI study had a substantially greater proportion of women than the historical studies. In general, the assessment of the historical data should identify differences in the outcomes for important subgroups so that the statistical analysis plan

for the NI study can be designed prospectively to take these factors into account (e.g., through covariate adjustments) or so that the value of M_1 can be revisited in light of the baseline characteristics of the study population that was enrolled in the NI study.

C. Fixed Margin and Synthesis Methods

Conceptually, the conclusion of non-inferiority is a synthesis of information from the NI trial itself and the historical evidence that the comparator in that trial was effective. The historical trials assessed $C_h - P$, the effect of the active control compared to placebo. The NI trial assesses $T - C_n$, the effect of the test drug compared to active control. If the outcome for the active control is constant across studies, then $C_n = C_h$, and the sum $T - C_n + C_h - P = T - P$ represents the effect of the test drug compared to placebo. The indispensable purpose of a NI trial is to show with high confidence that this effect is positive. Additionally, NI trials are usually intended to estimate this effect, at least roughly, in comparison to the historical effect of the active control.

The historical studies will furnish an estimate of $C_h - P$ with an associated standard error SE_H. Similarly, the NI study will estimate $T - C_n$ with a standard error SE_N. The two estimates are independent, so that the standard error of the sum is given by

$$\sqrt{SE_H^2 + SE_N^2} .$$

Notwithstanding that the interpretation of an NI study is fundamentally a synthesis, we recommend a statistical method, the fixed-margin method, that treats the problem in two separate steps. This approach offers two practical advantages. First, it allows separation of the problem of calculating, justifying, and possibly adjusting the NI margins M_1 and M_2 from the problem of analyzing the NI study. Second, it is equivalent to applying a somewhat larger standard error to the synthesized estimate, which can be seen as an allowance for the possibility of minor deviations from constancy.

1. The Fixed Margin Approach

The fixed margin approach to analysis in an NI trial is well known (see, e.g., ICH E9, section 3.3.2). This approach relies upon the choice of a fixed NI margin that is prespecified at the design stage of the NI trial and used to determine the sample size needed to provide sufficient power for a test of the hypothesis of non-inferiority. Operationally, the fixed margin approach usually proceeds in the following manner.

First, the active control effect is estimated from an analysis of past placebo-controlled studies of the active control (see section IV.B.2.a-b). The goal of this analysis is to define the margin M_1, a fixed value based on the past effect of the active control that is intended to be no larger than the effect the active control is expected to have in the NI study. The variability of the treatment effects observed across past studies as well as the constancy assumption should be taken into account in choosing M_1.

The selection of M_2 is then based on clinical judgment regarding how much of the active comparator treatment effect needs to be retained to demonstrate sufficient benefit for drug approval. The exercise of clinical judgment for the determination of M_2 should be applied after the determination of M_1 has been made. All relevant studies of the active comparator and all randomized patients within these studies should generally be used in determining the margin M_1 because that provides a more reliable estimate of the active control effect and avoids any potential for prejudice in selecting which historical studies to include. Determination of the relevance of past studies, however, may not be straightforward, and Examples 1(A) and 2 in the Appendix illustrate this point.

The lower bound of the confidence interval of the estimated active control effect based on past studies is typically selected as M_1. Concerns about past variability and constancy may lead to a determination to discount this lower bound in choosing M_1 to account for any sources of uncertainty and dissimilarities between the historical data and the NI study to be conducted, as discussed in earlier sections. Following this, a clinical judgment is made as to how much of this effect should be preserved. Choosing M_2 as 50% of M_1 has become usual practice for cardiovascular (CV) outcome studies, where effects may be small, and it is important to retain at least half of the effect of the original treatment. In antibiotic trials, effects tend to be quite large relative to placebo or ineffective treatments, and as a result, a 10-15% NI margin for the treatment difference (M_2) is commonly chosen. Note that the M_2 of 50% of M_1 is on a relative scale, whereas the 10-15% is on the risk difference scale for antibiotic drugs.

The formal test of the NI hypothesis also involves the use of a 95% confidence interval (i.e., to rule out $C\text{-}T > M_2$). Thus, there are two confidence intervals involved in the fixed margin approach, one for the historical data, where the lower bound is selected as M_1, and one for the NI study (to rule out $C\text{-}T > M_2$); both intervals are typically 95% confidence intervals. That is why this fixed margin approach is sometimes called the 95%-95% method.

Separating the process of estimating the treatment effect of the active comparator based upon the historical data (i.e., choice of M_1) from the analysis of the NI study has two advantages. It gives a single number that is clinically understandable for M_1 (and derived M_2) and provides a basis for planning the sample size of the NI study to achieve control of Type I error and the power needed for the NI study to meet its objective for the prespecified NI margin. Decisions to discount M_1 further or, where appropriate, to use a narrower confidence interval, are easily explained, and will make the fixed margin approach either more or less conservative.

Deciding on the NI clinical margin M_2 is a matter of judgment about how large a loss of the control treatment effect would be acceptable, a consideration that may reflect the seriousness of the outcome, the benefit of the active comparator, and the relative safety profiles of the test and comparator. Choice of M_2 also has major practical implications. For example, in large cardiovascular studies, it is unusual to have an M_2 that reflects a loss of less than 50% of the control drug effect, even if this might be clinically reasonable, because doing so will usually make the study size infeasible. Of course, allowing too much inferiority of the test drug to the standard, especially for endpoints of mortality and serious morbidity, would clearly not be acceptable.

The fixed margin approach considers the NI margin as a single number, fixed in advance of the NI study. The hypothesis tested in the NI study determines whether the comparison of the test drug to the active control meets the specified NI criterion, assuming, of course, that the active control had at least its expected effect (equal to M_1) (that is, the study had assay sensitivity). A successful NI conclusion, ruling out a difference > M_1, shows that the test drug is effective (just as a superiority study showing a significant effect at $p \leq 0.05$ does) and, if a difference > M_2 is also ruled out, shows that the new drug preserves the desired fraction of the control drug's effect.

2. The Synthesis Approach

An alternative statistical approach is known as the synthesis approach because it combines or synthesizes the data from the historical trials and the current NI trial, reflecting the variability in the two sources of data (the current NI study and the past studies used to determine HESDE). The synthesis method is designed to directly address the question of whether the test product would have been superior to a placebo had a placebo been in the NI study and also to address the related question of what fraction of the active control's effect is maintained by the test product. In the synthesis approach, M_1 is not prespecified, but the percent of active control effect to be preserved is prespecified. Based on the observed outcome of the test drug vs. active control comparison in the NI study, an evaluation is made as to whether the test agent has demonstrated preservation of the clinically relevant effect of the active control, without determining M_1.

Although the synthesis approach combines the data from the historical trials into the comparison of the concurrent active control and the test drug in the NI study, a direct randomized concurrent comparison with a placebo is of course not possible, because the placebo group is not a concurrent control and there is no randomization to such a group within the NI study. The imputed comparison with a placebo group that is not part of the NI study thus rests on the validity of several assumptions, just as the fixed margin approach does. The critical assumption of the constancy of the active control effect derived from the historical placebo-controlled trials is just as important when the synthesis method is used as when the fixed margin method is used.

The use of the synthesis approach can lead to a more efficiently designed study (e.g., by allowing for a reduction in sample size or achieving greater power for a given sample size) than the fixed margin approach, provided the constancy assumption holds. The greater statistical efficiency of the synthesis approach derives from how this method deals with the standard error of the comparison of test product to active control. See Appendix, Example 1(B), for a comparison of the two methods and the variance calculations.

The synthesis approach does not specify a fixed NI margin. Rather, the method combines (or synthesizes) the estimate of the treatment effect relative to the control from the NI trial with the estimate of the control effect from a meta-analysis of historical trials. The synthesis process makes use of the variability from the NI trial and the historical trials and yields one confidence interval for testing the NI hypothesis that the treatment preserves a fixed fraction of the control effect, without actually specifying that control effect or a specific fixed NI margin based on the control effect. Clinical judgment is used to prespecify an acceptable fraction of the control therapy's effect that should be retained by the test drug, regardless of the magnitude of the

control effect. The disadvantage of the synthesis approach is that it is not possible to use clinical judgment to choose M_2, based on the magnitude of M_1, in advance of the NI trial.

D. Considerations for Selecting the Clinical Margin (M_2)

M_2 is the prespecified NI margin that is to be ruled out in an NI study. The determination of M_2 is based on clinical judgment and is usually calculated by taking a percentage or fraction of M_1. The clinical judgment in determining M_2 may take into account the actual disease incidence or prevalence and its impact on the practicality of sample sizes that would have to be accrued for a study, as noted above with respect to cardiovascular outcome studies. There can be flexibility in selecting the M_2 margin, choosing a wider margin, for example, when:

1. The primary endpoint does not involve an irreversible outcome such as death (in general, the M_2 margin will be more stringent when treatment failure results in an irreversible outcome)
2. The test product is associated with fewer serious adverse effects or better tolerability than other therapies already available
3. The test product has another advantage over available therapies that warrants use of a less stringent margin (M_2)

A more stringent choice for M_2 may be required, however, when the difference between the active comparator response rate and the untreated response rate is large, making it feasible to demonstrate retention of a larger fraction of M_1 without requiring an impractical sample size for the NI study.

The implication of failing to rule out inferiority relative to M_1 and M_2 differs. Failure to exclude inferiority relative to M_1 means there is no assurance of any effect. Just as it would be unusual to accept a placebo-controlled study as positive (i.e., a finding of superiority) with $p > 0.05$, it would be unusual to accept an NI study as positive (i.e., a finding of non-inferiority) where the upper bound of the 95% confidence interval was > M_1. On the other hand, failing to exclude M_2 by a small amount (e.g., ruling out a loss of 52% of M_1 rather than a prespecified loss of 50% of M_1) may be acceptable, as the small amount would not suggest the absence of an effect of the drug.

E. Estimating the Sample Size

It is important to plan the sample size for an NI clinical trial so that the trial will have adequate statistical power to conclude that the NI margin is ruled out if the test drug is truly non-inferior. At the protocol planning stage, using the fixed margin approach, the NI margin (M_2) will be specified, and the sample size will be based on the need to rule out that margin. Both the size of the margin and the estimated variance of the treatment effect (T vs C) will affect the sample size determination. For event-driven trials, if the event rate in the NI study is lower than anticipated, power to demonstrate non-inferiority will be reduced. It may be important, therefore, to monitor overall (blinded) event rates during the study and to adjust the sample size if the interim event rates are unexpectedly low, a practice that is common in superiority trials. There is one further consideration in sample size planning. If, in reality, the test drug is somewhat more effective

than the control, it will be easier to rule out any given NI margin than if the test drug is equivalent or slightly inferior to the control, and a smaller sample size could be used. A somewhat less effective test drug will, of course, require a larger sample size.

It may be difficult to adequately plan the sample size for any study, including an NI study, because of uncertainty about assumptions such as event rates or endpoint variability at the time of study planning. For this reason, adaptive study designs that can allow for the prospective re-estimation of a larger sample size at one or more planned interim looks may be considered.

F. Study Quality and Choice of Analysis Population

Traditionally, the primary analysis of a randomized clinical superiority trial follows the intention-to-treat (ITT) principle, namely, all randomized patients are analyzed according to the treatment to which they were randomized, including patients who leave the study prematurely. This approach is intended to avoid various biases associated with patients switching treatment or patients being excluded from the analysis because of protocol violations or attrition. Adhering to the ITT principle in superiority trials is generally considered conservative, in that poor study quality resulting in a large number of protocol violations will tend to bias the results towards the null hypothesis of no difference between treatments. The opposite is true for NI trials. Quality issues could result in treatment groups appearing similar (i.e., biasing the results towards the alternative hypothesis for NI trials), when, in fact, the test drug may be inferior, as mentioned in section III.D.3. Many problems that may cause a superiority trial to fail, such as non-adherence, misclassification of the primary endpoint, or attrition, can bias the results toward no treatment difference (success) and undermine the validity of the trial, creating apparent non-inferiority when the test drug is in fact inferior. Imputation of missing data under the inferiority null hypothesis is one possible approach to countering the bias due to attrition.

The best advice for conducting an NI study is to emphasize study quality at the planning stage and continuously monitor the trial during the conduct and analysis stages to minimize the problems described above. If the NI trial is open label, attention to quality is all the more important because it may be very difficult to prove, after the fact, that enrollment, assessment of endpoints, and other study procedures were conducted in an unbiased way.

G. Testing Non-Inferiority and Superiority in a Single Trial

In general, when there is only one primary endpoint and one dose of the test treatment, a trial that is planned to demonstrate non-inferiority may also be used to test for superiority without concern about inflating the Type I error rate. This sequential testing procedure has the Type I error rates for tests of both non-inferiority and superiority controlled at the 2.5% level. A study designed primarily to show superiority, however, would yield credible evidence of non-inferiority only if the study had the key features of a NI study (e.g., the NI margin must be prespecified, and assay sensitivity and HESDE must be established). An unplanned determination of non-inferiority following failure to show superiority, when the margin was not determined until results of the trial were known, would not be sufficient for demonstrating non-inferiority of the test drug.

When multiple endpoints or multiple doses of the test treatment are to be evaluated in an NI study that also includes a test for superiority, careful planning of the order in which hypothesis tests are to be carried out is needed. For example, will superiority of the primary endpoint be tested before or after non-inferiority of a key secondary endpoint? A decision tree can be helpful in determining whether adjustments for multiplicity are needed and to what tests they should be applied to ensure that adequate control of the Type I error rate is achieved for the trial. In general, it would not be appropriate to apply 95% confidence intervals for multiple tests of non-inferiority and superiority across multiple endpoints or multiple doses because of the potential to inflate the Type I error rate for the trial; use of larger intervals that reflect adjustment for multiple tests (e.g., 97.5% confidence intervals or higher) may be needed.

V. FREQUENTLY ASKED QUESTIONS AND GENERAL GUIDANCE

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

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 only in situations that meet the following three general criteria for assessing the persuasiveness of a historically controlled trial (see ICH E10).

- First, there should be a good estimate of the historical untreated response 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 possibly including observational studies.
- Third, the untreated and treated patients should be comparable.

Then, if the treated and untreated response rates are substantially different, it may be possible to determine a NI margin. 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 rates are substantially different and can be used to determine M_1. The clinically acceptable loss of this effect can then be determined for M_2. Some examples of margins determined in this way have been presented in guidance on antibiotic trials, e.g., trials in community-acquired bacterial pneumonia (see Example 4 in the Appendix). Identifying a margin is more difficult 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%, identifying the NI margin in this case as 15% would not be reasonable because such a small difference could easily be the result of a 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 discussed in detail in sections III and IV, there is a critical difference between demonstrating 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% confidence interval did not rule out a loss of M_1 or greater would be similar to accepting as evidence of effectiveness a superiority study whose estimated treatment effect was not significant at $p \leq 0.05$. M_2, in contrast, represents clinical judgment about the amount of the active control effect that must be retained. A typical value for M_2 is often 50% of M_1, at least partly because the sample sizes needed to retain a larger amount, e.g., 60% or more, of the active control effect become impractically large. In this case, there is a better argument for some degree of flexibility if the study did not quite rule out the M_2 margin; there might be reason to consider, for example, assurance of 48% retention as acceptable. We have also concluded that the more conservative fixed margin method should generally be used in ensuring that M_1 is ruled out, but that the synthesis method can be used to assess non-inferiority with respect to M_2. Of course, allowing the test drug to be inferior to the standard by too large an amount, especially for endpoints of mortality and serious morbidity, would not be acceptable.

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

Prior information can be incorporated into a statistical model or within a Bayesian framework to take into account such factors as evidence of effects in other related indications or on other endpoints. As discussed in section IV.B.2.b, a meta-analysis is often used to estimate the average effect of the active control for purposes of setting the NI margin, and in certain circumstances, trials from related indications or for other drugs in the same class may be included in the meta-analysis conducted for this purpose. Some methods of meta-analysis allow the down-weighting of less relevant studies or studies that are not randomized or controlled (e.g., observational studies), which can be particularly important if few placebo-controlled trials are available.

Bayesian methods that incorporate historical information from past active control studies through the use of prior distributions of model parameters provide an alternative approach to evaluating non-inferiority in the NI trial itself. Although discussed in the literature and used in other research settings, CDER and CBER have not had much experience to date in evaluating NI trials of new drugs or therapeutic biologics that make use of a Bayesian approach for design and analysis. If a sponsor is planning to conduct a Bayesian NI trial, early discussions with the Agency are advised.

If important covariates are distributed differently in the historical studies than in the current NI study, model-based approaches may be used to adjust for these covariates in the NI analysis. Such covariates should be identified prior to the NI trial, and the methods for covariate adjustment should be specified prospectively in the NI trial protocol. Applying

post-hoc adjustments developed at the time of analyzing the NI trial would not be appropriate.

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

The active control does not have to be labeled or approved in the United States for the indication being studied in the NI study as long as there are adequate data that are reliable and reproducible to support the chosen NI margin. FDA does, in some cases, rely on published literature and has done so in carrying out the meta-analyses of the effect of the active control that are used to define NI margins. The FDA guidance, "Providing Clinical Evidence of Effectiveness," describes the approach to considering the use of literature in providing evidence of effectiveness, and similar considerations would apply here. Among these considerations are:

- The quality of the publications (the level of detail provided)
- The difficulty of assessing the endpoints used
- Changes in practice between the present and the time of the studies
- Whether FDA has reviewed some or all of the studies
- Whether FDA and the sponsor have access to the original data

As noted above, the endpoint for the NI study could be different (e.g., death, heart attack, and stroke) from the primary endpoint (cardiovascular death) in the studies if the alternative endpoint is well assessed and data on that endpoint are available for determining M_1 (see also question 6).

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

When an active control drug is approved for the same indication as that of the NI trial, the size of the effect appearing in the label of the active control is usually based on the pivotal trials rather than a meta-analysis of all past studies. Further, the variability of the effect across past studies may not be available. In general, approval of a drug is based on showing superiority to placebo, usually in at least two studies, but FDA may not have critically assessed whether the effect is consistent across past studies and may not have analyzed any "failed" studies. It is therefore essential to use the data from all available controlled trials (unless a trial has a significant defect), including trials conducted after marketing, to calculate a reasonable estimate of the actual control effect. If the active control data have been used to define the NI margin for another study of the same indication, and if a determination is made that the trials involved are relevant to the new trial, then use of the same margin in the new trial should be acceptable.

6. What factors should be considered when selecting an endpoint for a NI trial?

The endpoints chosen for clinical trials (superiority or non-inferiority) should be clinically meaningful measures of the way patients feel, function, or survive that can be reliably assessed in the target population. The endpoints generally reflect the event rate or other measure of disease in the population but also must take into account practical considerations, such as the size of the study required to demonstrate superiority or non-inferiority with respect to the endpoint. In NI studies, the endpoint must be one for which there is a good basis for knowing the effect of the active control. The endpoint need not necessarily be the endpoint used in the historical trials or the effectiveness endpoint appearing in labeling of the active control. Past trials that were successful in showing an effect on a mortality endpoint could, for example, serve as the basis for estimating an effect on a composite endpoint (cardiovascular mortality, myocardial infarction, and stroke), if that were the desired endpoint for the NI study. The use of a different endpoint in the NI study might be desirable because it would permit a smaller study, but it would be important not to include components that were not affected by the active control or did not represent an important clinical benefit.

7. Are there circumstances where conducting an NI study may not be feasible?

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

- The treatment effect of the active comparator may be so small that the sample size required to conduct a NI study may not be feasible (unless the test drug is assumed to be superior).
- There is large study-to-study variability in the treatment effect. In this case, the treatment effect may not be sufficiently reproducible, calling into question the assumption of assay sensitivity. This is often the case for symptomatic treatments (e.g., treatments for depression, anxiety, insomnia, angina, symptomatic heart failure, symptoms of irritable bowel disease, and pain).
- There is no historical evidence available to determine a NI margin, and assuming a zero response rate for untreated patients is not reasonable.
- Medical practice has changed so much (e.g., the active control is always used with additional drugs) that the effect of the active control in the historical studies is not relevant to the current study.

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

In that case it may be possible to design a superiority study in an appropriate population that would be considered ethical. Several possibilities are discussed in ICH E10 and include the following:

- When the new drug and an established treatment are pharmacologically distinct, an add-on study where the test drug and placebo are each added to the established treatment.
- A study in patients who do not respond to the established therapy. It may be possible to conduct a placebo-controlled trial in these patients. Alternatively, nonresponders could be randomly assigned to the test drug or the failed therapy, and superiority assessed to demonstrate effectiveness in the nonresponder population.
- A study in patients who cannot tolerate the established effective therapy.
- A study of a population for which there is no available effective therapy.
- A dose-response study without placebo, if the new drug is known to have dose-related side effects, and a dose lower than the usual dose would be considered ethical.

9. When can a single NI study be sufficient to support effectiveness?

Several sections above touch on this question, notably section III.H, which discusses this issue in detail. Briefly, reliance on a single study in the NI setting is based on considerations similar to reliance on a single study in the superiority setting, with the additional consideration of the stringency of showing non-inferiority using the M_2 margin. Many of these considerations are described in the FDA guidance, "Providing Clinical Evidence of Effectiveness", and include supportive information, such as results with pharmacologically similar agents (a very common consideration, because the NI study will often compare drugs of the same pharmacologic class), support from credible biomarker information (tumor responses, ACE inhibition, beta blockade), and a statistically persuasive result from the single NI study. With respect to the latter, it is noted above that a finding of NI based on excluding a treatment difference > M_2 provides very strong evidence that the test treatment has an effect > 0 when M_2 is substantially smaller than M_1. For all these reasons, most NI studies with mortality or serious morbidity as endpoints, if clearly successful, will be sufficient to demonstrate NI as single studies.

APPENDIX — EXAMPLES

The following four examples derived from publicly available information (see references following examples) illustrate determination of the NI margin, application of methods of NI analysis, and other considerations relevant to determining whether it is possible to conduct and interpret the results of an NI study.

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

This example will demonstrate the following:

- 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 control effect
- How to assess whether the assumption of assay sensitivity is appropriate and whether the constancy assumption is reasonable for this drug class
- The use of 95% confidence intervals for both margin determination and in the NI study for C-T to demonstrate non-inferiority, i.e., the 95% - 95% method

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 nonvalvular atrial fibrillation at risk of thromboembolic complications (e.g., stroke, TIA). There are six placebo-controlled studies of warfarin involving the treatment of patients with nonvalvular 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 (ischemic and hemorrhagic strokes plus systemic embolic events) on warfarin compared to placebo is favorable to warfarin in each of the six 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 significant treatment effect in favor of warfarin. 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, which might have introduced a favorable bias. These placebo-controlled studies of warfarin in patients with nonvalvular atrial fibrillation show a fairly consistent effect. Based on the positive results from five of the six studies, it can reasonably be assumed that were placebo to be included in a warfarin-controlled NI study involving a novel anticoagulant, warfarin would have been superior to placebo.

Table 1. Placebo-Controlled Trials of Warfarin in Nonvalvular 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)
EAFT	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)

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

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 was judgment involved in assessing the endpoints, there may have been the potential for bias to be introduced. The event rate for placebo in the EAFT study was strikingly high, perhaps because the patient population in that study was different from the patient population in the remaining five studies, in that only patients with a recent TIA or stroke were enrolled in EAFT. That factor would clearly increase the event rate, but in fact the risk reduction in EAFT 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.

Even if the historical studies are consistent, a critical consideration in deciding upon the NI margin is whether the constancy assumption is reasonable. That is, we must consider whether the magnitude of the 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 the desired International Normalised Ratio (INR)), or changes in the population being studied may make the effect of warfarin estimated from the previous studies not relevant to the current NI study.

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 characteristics can be compared across the studies (e.g., age, race, and target INR), some cannot (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 the question. Table 2 shows that for some characteristics, such as a history of stroke or TIA, there are some differences across the studies. An important inclusion criterion in the EAFT study was that subjects have a prior history of stroke or TIA. None of the other studies had such a requirement. Subjects enrolled into the EAFT study were thus at higher risk than subjects in the other studies, presumably leading to the higher event rates in both the warfarin and placebo arms, shown in Table 1. The higher event rates in the EAFT study may also have been influenced by the relatively long duration of follow-up or the fact that the primary endpoint definition was broader (including vascular deaths and nonfatal myocardial infarctions), or both. Even with the

higher event rates observed in this one study, however, the risk ratios are quite consistent (with the exception of CAFA), a relatively reassuring outcome.

Table 2. Demographic Variables, Clinical Characteristics, and Endpoints of Warfarin Atrial Fibrillation 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

* = 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

To calculate M_1, the relative risks in each of the six studies were combined using a random effects meta-analysis to give a point estimate of 0.361 for the relative risk with a confidence interval of (0.248, 0.527). The 95% confidence interval 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.

It was considered clinically important to show that the test drug preserved a substantial fraction of the warfarin effect. The clinical margin M_2 representing the largest acceptable level of inferiority was therefore set at 50% of M_1. M_2 was calculated on the log hazard scale as 1.378, so that NI would be demonstrated provided the upper bound for the 95% confidence interval for C vs. T < 1.378.

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

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

This example demonstrates the following:

- The critical features of the synthesis method for demonstrating non-inferiority of a new anticoagulant
- The calculations and sources of statistical variability that are incorporated in the synthesis approach
- The main differences in interpretation of results from the fixed margin and the synthesis approaches when applied to the same set of studies and data

In this example, we illustrate the synthesis method using the same data as Example 1(A), which consist of six studies comparing warfarin to placebo and one NI study comparing ximelegatran to warfarin. In contrast to the fixed margin method in Example 1(A), the synthesis method does not use a separate 95% confidence interval for this historical estimate of the effect of warfarin versus placebo and for the comparison in the NI study. Rather, the synthesis method is constructed to address the questions of whether ximelegatran preserves a specified percent, in this case 50% of the effect of warfarin, and whether ximelegatran would be superior to a placebo, if placebo had been included as a randomized treatment group in the NI study. To accomplish this goal, the synthesis method makes a comparison of the effect of ximelegatran in the NI study to historical placebo data, an indirect comparison that is not based upon a randomized concurrent placebo group. The synthesis method combines the data from the placebo-controlled studies of warfarin with the data from the NI study in such a way that a test of hypothesis is made to demonstrate that a certain percent of the effect of warfarin is retained in the NI study. A critical point distinguishing the synthesis method from the fixed margin method is that the effect (M_1) of warfarin is not specified in advance. But to carry out the analysis, an assumption needs to be made regarding the placebo comparison, namely, that the difference between control drug and placebo (had there been one) in the NI trial is the same as that seen in the historical placebo-controlled trials of warfarin. The assumption is needed because there is no randomized comparison of warfarin and placebo in the NI trial. As a point of reference, we know from example 1(A) that the warfarin effect M_1 was estimated from the historical placebo-controlled studies to be a 47% risk reduction.

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

H_0: {log-relative risk of ximelegatran versus warfarin} ≥

– ½ {log-relative risk of warfarin versus placebo}

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-relative risk of ximelegatran versus warfarin} + ½ {log-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 .361) with a standard error of 0.154. These estimates are combined, and the synthesis test statistic is calculated as:

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

Assuming the statistic is normally distributed, it is then compared to − 1.96 (for a one-sided Type I 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 using the synthesis method.

The fixed margin test statistic for this example is:

$$\frac{0.329 + \frac{1}{2}(-1.02)}{0.216 + \frac{1}{2}(0.154)} = -0.618$$

This value, -0.618, is also greater than -1.96. Neither method leads to a conclusion of non-inferiority for this example.

Example 2: Aspirin to Prevent Death or Death/MI After Myocardial Infarction

This example demonstrates the following:

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

By 1993, the effect of aspirin in preventing death after myocardial infarction had been studied in six large randomized placebo-controlled clinical trials. A seventh trial, ISIS-2, gave the drug during the first day after the AMI and is not included because it addressed a different question. The results are summarized and presented in Table 3.

Table 3. Results of Six Placebo-Controlled Randomized Studies of the Effect 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)

The results suggest:

1. The effect of aspirin on mortality as measured by the relative risk appears to diminish over the years in which the studies were conducted.
2. The largest trial, AMIS, showed a numerically higher mortality rate in the aspirin-treated group.

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

Although a fixed effects meta-analysis of the six studies gives a point estimate of 0.91 (95% confidence interval 0.82 to 1.02), the random-effects analysis gives a point estimate of 0.86 with 95% confidence interval (0.69, 1.08). The effect of aspirin on prevention of death after myocardial infarction in these historical studies is thus inconclusive (i.e., the upper bound of the 95% confidence interval for effect is > 1.0). Therefore, it is not possible to select aspirin as the active control for evaluating the mortality effect of a test drug in an NI trial. Even if the upper

bound excluded 1.0, it would be difficult to rely on an NI margin that is not supported by AMIS, the largest of the six trials.

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

Table 4. Results of Six Placebo-Controlled Randomized Studies of the Effect of Aspirin in Secondary 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)

*The event rates of both groups need further verification from each article.

The results indicate that the effect of aspirin on death or MI after myocardial infarction is small to absent in the latest trial (AMIS). Random effects meta-analyses give point estimates of the relative risk of approximately 0.8 with 95% confidence interval upper bounds of approximately 1.0. The NI margin based on these six studies is close to zero (without reducing it further to represent M_2) and is so small that an NI trial would be unrealistically large. Again, as with the mortality endpoint, it would be troubling even to consider an NI approach when the largest and most recent trial showed no significant effect.

Example 3: Xeloda to Treat Metastatic Colorectal Cancer — the Synthesis Method

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

- The use of the synthesis method to demonstrate a loss of no more than 50% of the historical control treatment's effect and a relaxation of this criterion when two NI studies are available
- The use of additional endpoints in the decision-making process
- The use of a conservative estimate of the active control effect because (1) a subset of the available studies to estimate the margin was selected and (2) the effect was measured relative to a previous standard of care instead of placebo

The U.S. regulatory standard for first-line treatment of metastatic colorectal cancer, the use sought for Xeloda, is the demonstration of improvement in overall survival. Two separate clinical trials, each using an NI study design, compared Xeloda to a Mayo Clinic regimen of 5-fluorouracil with leucovorin (5-FU+LV), the standard of care at the time. Xeloda is an oral fluoropyrimidine, while 5-fluorouracil (5-FU) is an infusional fluoropyrimidine.

By itself, bolus 5-FU had not demonstrated a survival advantage in first-line treatment of metastatic colorectal cancer. But with the addition of leucovorin to bolus 5-FU, the combination had demonstrated improved survival. A systematic evaluation of approximately 30 studies that investigated the effect of adding leucovorin to a regimen of 5-FU identified 10 clinical trials that compared a regimen of 5-FU+LV similar to the Mayo Clinic regimen to 5-FU alone, thereby providing a measure of the effect of LV added to 5-FU — a conservative estimate of the overall effect of 5-FU+LV as it is likely 5-FU has some effect.

Table 5 summarizes the overall survival results, using the metric "log hazard ratio" for the 10 studies identified that addressed the comparison of interest.

Table 5. Selected Studies Comparing 5-FU 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

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

A random effects model applied to the survival results of these 10 studies yielded the historical estimate of the 5-FU versus 5-FU+LV survival comparison of a hazard ratio of 1.264 with a 95%

confidence interval of (1.09, 1.46) and a log hazard ratio of 0.234. The NI margin is therefore 1.09 for a fixed margin approach ruling out M_1.

A summary of the survival results based on the intent-to-treat populations for each of the two Xeloda NI trials is presented in Table 6. Study 2 rules out M_1 using a fixed margin approach, but Study 1 does not.

Table 6. Summary of the Survival Results

Study	Hazard Ratio[1]	Log Hazard Ratio[1]	Standard Error	95% confidence interval 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)

[1] Hazard ratios and log hazard ratios are Xeloda/5-FU+LV

The clinical choice of how much of the effect on survival of 5-FU+LV should be retained by Xeloda was determined to be 50%. The synthesis approach was used to analyze whether the NI criteria of 50% loss was met. This synthesis approach for each study combines the results of that NI study with the results from the random effects meta-analysis to produce a normalized test statistic.

Based on this NI synthesis test procedure, NI Study 1 failed to demonstrate that Xeloda retained at least 50% of the historical effect of 5-FU+LV versus 5-FU on overall survival, but NI Study 2 did demonstrate such an effect. It was then decided to determine what percent retention was demonstrated. By adapting the synthesis test procedure for retention of an arbitrary percent of the 5-FU+LV historical effect, it was determined that NI Study 1 demonstrated that Xeloda lost no more than 90% of the historical effect of 5-FU+LV on overall survival and that NI Study 2 demonstrated no more than a 39% loss of the historical effect.

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

Example 4: Determination of an NI margin for Community-Acquired Bacterial Pneumonia (CABP) When No Historical Trials Are Available

This example illustrates the following points:

- The use of risk differences as the metric for estimated treatment effects
- The determination of the NI margin using observational data when no randomized, placebo-controlled trials of active control drugs are available

Community-acquired bacterial pneumonia (CABP) is an acute bacterial infection of the pulmonary parenchyma associated with chest pain, cough, sputum production, difficulty breathing, chills, rigors, fever, or hypotension, and is accompanied by the presence of a new lobar or multi-lobar infiltrate on a chest radiograph. FDA issued revised draft guidance for developing treatments of CAPB in 2014, and the appendix in that guidance describes in detail the justification for NI margins with respect to two endpoints: clinical response and mortality.[3] Historical data from nonrandomized studies of bacteremic and nonbacteremic patients with pneumococcal or lobar pneumonia were evaluated to justify NI margins for use in future CABP studies. For the clinical response endpoint, historical data are relied upon not only to determine an acceptable NI margin but also to decide the time point following treatment at which non-inferiority should be assessed. For the mortality endpoint, margin determination is more straightforward and provided here for illustration.

An area of uncertainty in evaluating historical data is the spectrum of bacterial pathogens that cause CABP today. In most of the historical studies, CABP was considered synonymous with pneumococcal pneumonia because *S. pneumoniae* was regularly identified. CABP is also caused by other pathogens such as *H. influenzae*, *H. parainfluenzae*, *S. aureus*, and *M. catarrhalis*, as well as atypical bacteria such as *M. pneumoniae*, *C. pneumoniae*, and *Legionella* species. A fundamental assumption is that historical response rates in infections such as *S. pneumoniae* CABP are relevant to response rates in modern infections with sensitive organisms.

Table 7 provides an overview of the historical studies available for estimation of mortality rates among both treated and untreated patients with pneumococcal pneumonia. In all of the studies, a significant mortality benefit was shown for patients treated with antibacterial drugs (including sulfonamides, penicillin, and tetracyclines) compared to patients who received no specific therapy (untreated).

[3] See the revised draft guidance for industry *Community-Acquired Bacterial Pneumonia: Developing Drugs for Treatment*. When final, this guidance will represent the FDA's current thinking on this topic.

Table 7. Mortality in Observational Studies of Pneumococcal Pneumonia

Publication	Population	Mortality (%): Untreated Patients (Study Years)	Mortality (%): Antibacterial-Treated Patients (Study Years)	Treatment Difference Untreated-Treated (95% CI)
Finland (1943)	≥ 12 years old bacteremic and nonbacteremic	N=2,832 (1929-1940)* 41%	N=1,220 (1939-1941) 17% (sulfonamides)	24% (21%, 27%)
Dowling and Lepper (1951)	≥ 10 years old bacteremic and nonbacteremic	N=1,087 (1939, 1940)* 30.5%	N=1,274 (1938-1950) 12.3% (sulfonamides) N=920 (1938-1950) 5.1% (penicillins and tetracyclines)	18.2% (15%, 21%) 25.4% (22%, 28%)
Austrian and Gold (1964)	≥ 12 years old bacteremic	N=17 (1952-1962) 82%	N=437 (1952-1962) 17%	65% (41%, 79%)

* Historical control patients

There are two limitations in the use of these data to determine an NI margin. First, only observational data are presented, and second, no recent studies were available, with the most recent dating to the 1960s. Significant improvements in the standard of care and the availability of better treatment options for patients with CABP compared to the preantibiotic era bring into question the relevance of these historical studies in estimating an active control effect with respect to mortality. In spite of these limitations, the mortality rates among the treated patient cohorts are reasonably consistent, ranging from 5% to 17% across the three decades represented. Mortality rates in the untreated cohorts are more variable. The older references provide estimates of 31% and 41%, while the more recent study gives an estimated mortality rate of 82%, but this estimate is based on a very small sample size (n=17).

In the absence of any placebo-controlled trials of active control drugs in this disease area, it is not possible to estimate M_1 using the methods advocated in this guidance document (e.g., the 95%-95% method), but it is clear that the untreated mortality is substantial. Thus, based on the historical evidence described above, with the caveats noted about the nature and age of the studies, it is reasonable to assume that the mortality rate due to CABP, if left untreated, will be substantially higher than the rates observed among the treated cohorts. Use of an NI margin (M_2) of approximately 10% is therefore a valid approach for evaluating new treatments of CABP and would clearly represent an effect superior to no treatment as well as, based on clinical judgment, an appropriate clinical margin.

REFERENCES FOR EXAMPLES

Example 1(A and B)

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, 431-438.

Jackson, K., Gersh, B.J., Stockbridge, N., Fleming, 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.

Example 2

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

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

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

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

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

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

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

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

Example 3

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

FDA Medical-Statistical Review for Xeloda (NDA 20-896) dated April 23, 2001. http://www.accessdata.fda.gov/drugsatfda_docs/nda/2001/20896s6_Xeloda_Medr_Statr_P2.pdf.

Example 4

FDA Draft Guidance for Industry (Revision 2): *Community-Acquired Bacterial Pneumonia: Developing Antimicrobial Drugs for Treatment*, January 2014.

Singer, M., Nambiar, S., Valappil, T., Higgins, K., and Gitterman, S. (2008). "Historical and Regulatory Perspectives on the Treatment Effect of Antibacterial Drugs for Community-Acquired Pneumonia." *Clin Infect Dis* 47 (Suppl 3), S216-S224.

Finland, M. (1943). "Chemotherapy in the Bacteremia." *Conn State Med J* 7, 92-100.

Dowling, H.G., and Lepper, M.H. (1951). "The Effect of Antibiotics (Penicillin, Aureomycin and Terramycin) on the Fatality Rate and Incidence of Complications in Pneumococcic Pneumonia: A Comparison With Other Methods of Therapy." *AM J Med Sci* 222, 396-402.

Austrian, R., and Gold, J. (1964). "Pneumococcal Bacteremia With Especial Reference to Bacteremic Pneumococcal Pneumonia." *Ann Intern Med* 60, 759-776.

REFERENCES

General

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.

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., Campbell, M., Greenhouse, J. (2005). "Non-Inferiority Trials: Continued Advancements in Concepts and Methodology." *Statistics in Medicine* 25, 1097-1099.

Fleming, T.R. (2008). "Current Issues in Non-inferiority Trials." *Statistics in Medicine* 27, 317-332.

Fleming, T.R. (2011). "Some Essential Considerations in the Design and Conduct of Non-inferiority Trials." *Clinical Trials* 8, 432-439.

Hauschke, D., Pigeot, I. (2005). "Establishing Efficacy of a New Experimental Treatment in the 'Gold Standard' Design (with discussions)." *Biometrical Journal* 47, 782-798.

Koch, G.G. (2008). "Comments on 'Current Issues in Non-Inferiority Trials.'" *Statistics in Medicine* 27, 333-342.

Peterson, P., Carroll, K., Chuang-Stein, C., Ho, Y-Y., Jiang, Q., Gang, L., Sanchez, M., Sax, R., Wang, Y-C., Snapinn, S. (2010). "PISC Expert Team White Paper: Toward a Consistent Standard of Evidence When Evaluating the Efficacy of an Experimental Treatment From a Randomized Active-Controlled Trial." *Statistics in Biopharmaceutical Research* 2, 522-531.

Rothmann, M., Li, N., Chen, G., Chi, G.Y.H., Temple, R.T., Tsou, H.H. (2003). "Non-inferiority Methods for Mortality Trials." *Statistics in Medicine* 22, 239-264.

Snapinn, S.M., Jiang, Q. (2013). "Remaining Challenges in Assessing Non-inferiority." *Therapeutic Innovation & Regulatory Science* 48(1), 62-67.

Regulatory

Committee for Proprietary Medicinal Products (CPMP)(2001). Points to Consider on Switching between Superiority and Non-Inferiority. *Br J Clin Pharmacol* 52, 223-228.

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

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.

Huitfeldt, B., Hummel, J., on behalf of European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) (2011). "The Draft FDA Guideline on Non-Inferiority Clinical Trials: A Critical Review From European Pharmaceutical Industry Statisticians." *Pharmaceutical Statistics* 10, 414-419.

Hung, H.M.J., Wang, S.J., O'Neill, R.T. (2005). "A Regulatory Perspective on Choice of Margin and Statistical Inference Issue in Non-Inferiority Trials." *Biometrical Journal* 47, 28-36.

International Conference on Harmonisation: *Statistical Principles for Clinical Trials* (ICH E-9), Food and Drug Administration, DHHS, 1998.

International Conference on Harmonisation: *Choice of Control Group and Related Issues in Clinical Trials* (ICH E10), Food and Drug Administration, DHHS, July 2000.

Temple, R., Ellenberg, S.S. (2000). "Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments - Part 1: Ethical and Scientific Issues." *Annals of Internal Medicine* 133, 455-463.

Synthesis Methods

Brittain, E.H., Fay, M.P., Follmann, D.A. (2012). A valid formulation of the analysis of non-inferiority trials under random effects meta-analysis. *Biostatistics* 13, 637-649.

Hasselblad, V., Kong, D.F. (2001). "Statistical Methods for Comparison to Placebo in Active-Control Trials." *Drug Information Journal* 35, 435-449.

Rothmann, M.D., Wiens, B.L., and Chan, S.F. (2012). *Design and Analysis of Non-Inferiority Trials (Chapter 5.3).* Boca Raton, FL:Taylor and Francis Group.

Meta-analysis

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

Follmann, D.A., Proschan, M.A. (1999). "Validity Inference in Random-Effects Meta-Analysis." *Biometrics* 55, 732-737.

Muthukumarna, S. and Tiwari, R.C. (2012). "Meta-analysis Using Dirichlet Process." *Statistical Methods in Medical Research* D0I: 10.1177/0962280212453891.

Bayesian Approaches

Gamalo, M., Tiwari, R.C., LaVange, L.M. (2014). "Bayesian Approach to the Design and Analysis of Non-inferiority Trials for Anti-infective Products." *Pharmaceutical Statistics* 13(1), 25-40.

Simon, R. (1999). "Bayesian Design and Analysis of Active Control Clinical Trials." *Biometrics* 55, 484-487.

Study Quality and Analysis Population

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.

Sanchez, M.M., Chen, X. (2006). "Choosing the Analysis Population in Non-Inferiority Studies: Per Protocol or Intent-to-Treat." *Statistics in Medicine* 25, 1169-1181.

In-Text Examples

Retavase (reteplase) label available on the Internet at: https://dailymed.nlm.nih.gov/dailymed/archives/fdaDrugInfo.cfm?archiveid=13948

Scandinavian Simvastatin Survival Study Group (1994). "Randomized Trial of Cholesterol Lowering in 4444 Patients with Coronary Heart Disease: The Scandinavian Simvastatin Survival Study (4S)." *Lancet* 344,1383-1389.

Shepherd J., Cobbe S.M., Ford L., Isles C.G., Lorimer A.R., Macfarlane P.W., McKillop J.H., Packard C.J., for the West of Scotland Coronary Prevention Group (1995). "Prevention of Coronary Heart Disease with Pravastatin in Men with Hypercholesterolemia." *N Engl J Med* 333,1301-7.

Downs J.R., Clearfield M., Weis S.,Whitney E., Shapiro D.R., Beere P.A., Langendorfer A., Stein E.A., Kruyer W., Gotto A.M., for the AFCAPS/TexCAPS Research Group (1998). "Primary Prevention of Acute Coronary Events with Lovastatin in Men and Women with Average Cholesterol Levels." *JAMA* 279(20), 1615-1622.