

Non-inferiority Designs: Concepts and Methods

James Ware

Victor De Gruttola

Harvard School of Public Health

Background

Recognizing that patients should not be denied effective treatment, investigators have increasingly turned to trials comparing experimental therapy to an active control, that is, an effective treatment. This has led to the development of the **non-inferiority design**.

In the NI design, the experimental treatment, T, is compared to an active treatment, C, that has previously been shown to be effective, that is, superior to placebo, in the study population and treatment setting under investigation.

In the NI trial, the goal is to show that the new treatment, T, is **not inferior** to C.

Superiority Trials

In the superiority trial, we compare an investigational drug to standard therapy or a placebo.

The goal is to demonstrate the superiority of the new treatment to the placebo (no treatment) or the standard regimen.

To do so, we specify a **null** hypothesis, that the two treatments are equally effective with respect to a specified primary endpoint, and an **alternative** hypothesis, that the new treatment is superior.

We hope that the data will support the decision to reject the null hypothesis and lead to the conclusion that the alternative hypothesis is true, namely, that the new treatment is superior to standard care.

Hypothesis Testing in the Superiority Trial

To be specific, assume that the outcome is a binary event, for example, death, in a specified time period. We call this a **binary** endpoint because there are two possibilities, the event either happens or does not happen.

Let p_P be the probability that a study participant in the placebo group will experience the event and p_T be the probability of the event in the active treatment group. Then the study hypotheses are:

H_0 : $p_T = p_P$, (equally effective) and

H_A : $p_T < p_P$ (T superior to P)

Study Design for a Superiority Trial

When the trial is completed, we will use standard statistical methods to compare the event rates in the two groups. We will make one of two decisions:

If the event rates are significantly different, we will **reject** the null hypothesis.

If the event rates are not significantly different, we will **fail to reject the null hypothesis**.

Note that we can never prove the null hypothesis, in this case that the mortality rates in the two groups are exactly equal. We only conclude that the evidence does not disprove the null hypothesis.

Possible Outcomes of Hypothesis Testing

Unobserved Truth in the Population





H_A : T is superior
to P

H_0 : T and P are
equally effective

**Decision
Based on
Statistical
Analysis**

Reject H_0 :
T superior
to P

Fail to
reject H_0 : T
not shown
superior

True positive 	Type I error 
Type II error 	True negative 

Superiority Trials

The details of the superiority design vary according to whether the primary outcome is binary, a measurement, or time to an event. However, the general approach is the same for all of these settings.

We turn now to a discussion of the principles and issues that arise in designing and interpreting non-inferiority trials. We will not delve deeply into a detailed discussion of statistical methods, but will focus instead on the concepts.

Non-inferiority Designs

In the non-inferiority trial, we compare a new treatment (T) to a comparator treatment (C) that has previously been shown to be effective in the treatment of the target condition.

Since C is effective, we need not show that T is superior to C to demonstrate clinical benefit of T. Instead, we must show that it is **non-inferior**. Implicitly, by showing that T is not inferior to C, we show that T is superior to P. This is called the **implicit placebo comparison**.

The non-inferiority design can be appropriate when:

- T has fewer side effects or is less costly

- T has other benefits

- T is a “me-too” drug, a member of a class of drugs known to be effective, e.g., NRTI’s.

Non-inferiority Designs

suppose that we wish to design a non-inferiority trial whose endpoint is a composite of death and virological failure.

Let p_C be the true event rate in the comparator arm and p_T be the event rate in the new treatment arm.

Recall that, to conduct a superiority trial, we would specify the null and alternative hypotheses as follows:

$$H_0: p_C = p_T$$

$$H_A: p_T < p_C$$

Non-inferiority Designs

In the non-inferiority trial, we would like to show that T is **as good as** C. It is impossible to show that two treatments have identical efficacy. Instead, in the non-inferiority design, we choose a non-inferiority margin, M , and seek to prove $p_T - p_C$ is less than M . This is done by testing the null hypothesis:

$$H_0: p_T - p_C \geq M \quad (\text{The event rate in the T group is larger by } M \text{ or more})$$





against the one-sided alternative

$$H_A: p_T - p_C < M \quad (\text{If T is less effective, the difference in event rates is less than } M.)$$

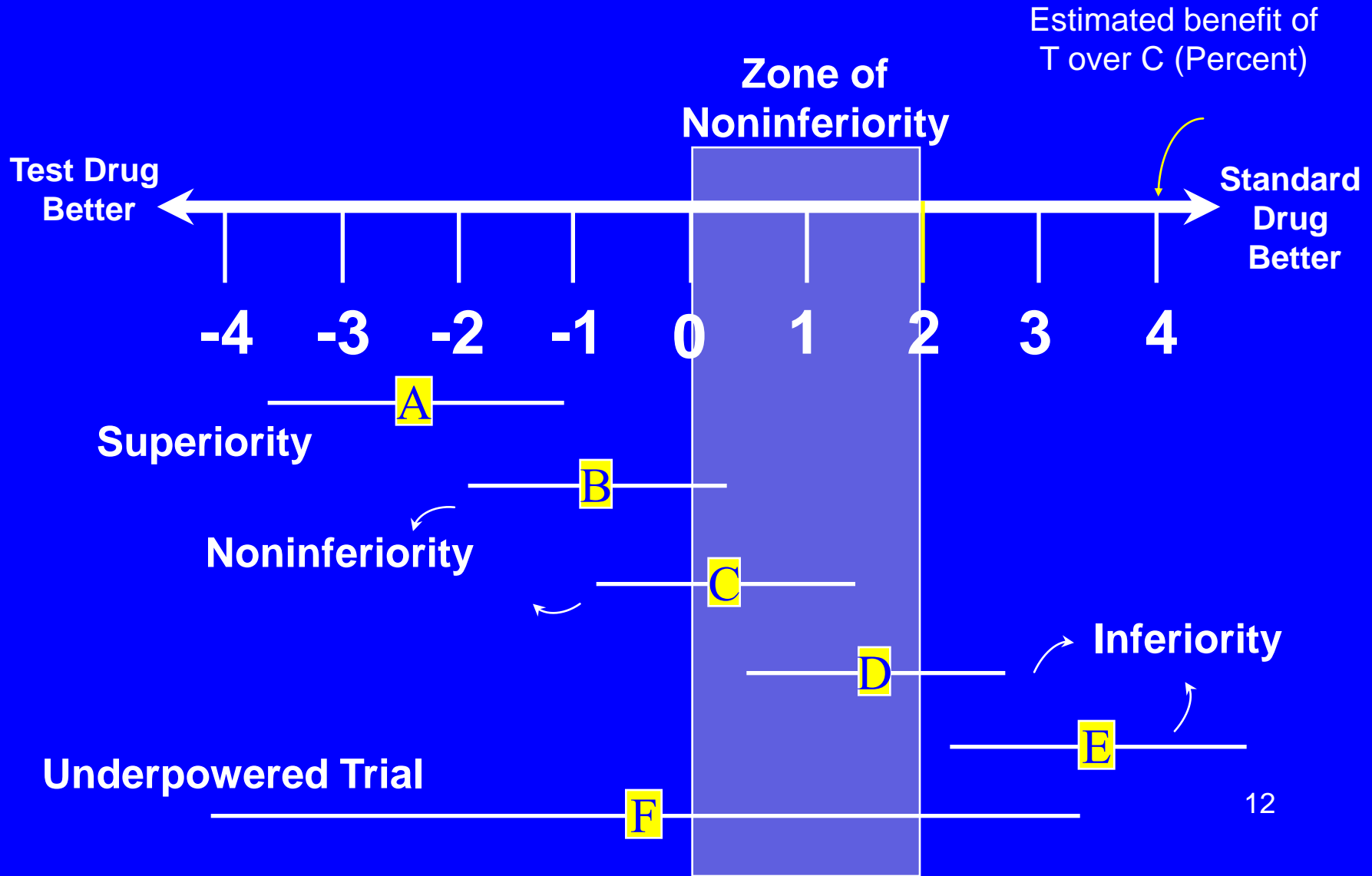
If we can reject H_0 , we can infer that T increases the event rate by less than M compared to C.

Possible Outcomes in a Non-Inferiority Trial

Unobserved Truth in the Population

		H_A : T not inf to C	H_0 : T is inferior to C
Decision Based on Statistical Analysis	Reject H_0 : T not inferior to C	True positive 	Type I error 
	Fail to reject H_0 : T not shown non- inferior	Type II error 	True negative 

Difference in Events : Test Drug - Standard Drug



Example: The ACUITY Trial

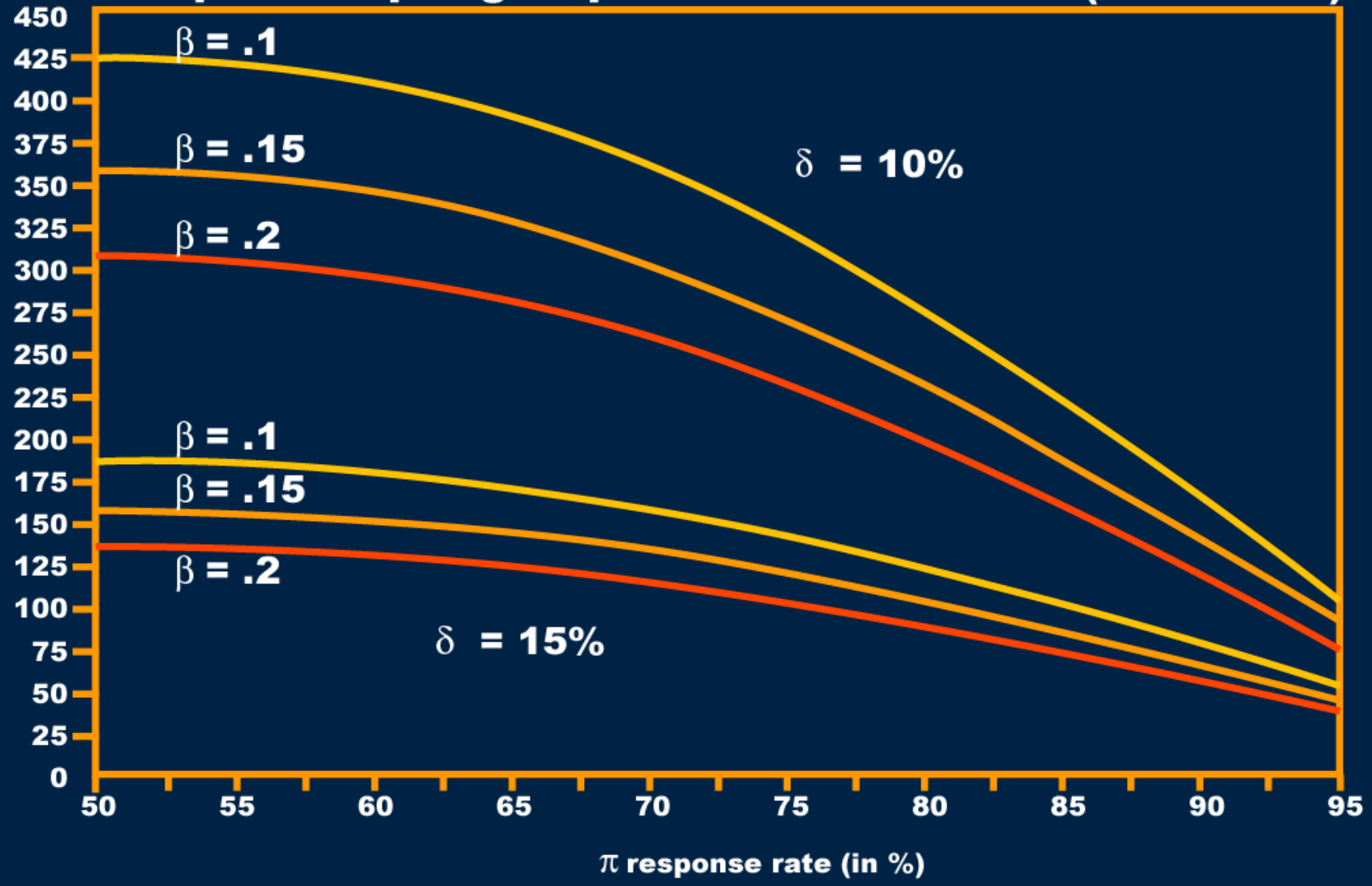
The Medicines Company sponsored the ACUITY trial³ to assess the efficacy of their product, bivalirudin, relative to heparin for the treatment of acute coronary syndrome.

Since heparin was known to be effective in reducing risk of ischemic events for such patients, the investigators chose a non-inferiority design.

The primary endpoint was death, MI, or unplanned revascularization within 30 days of enrollment.

sample size per group

$\alpha = .05$ (one-sided)



The ACUITY Trial

The investigators anticipated a 30-day event rate of 6.5% in the heparin group. They chose a non-inferiority margin corresponding to a 25% increase in the event rate. Expressed as an absolute increase in the event rate, the non-inferiority margin was

$$6.5 \times 0.25 = 1.625\%$$

Thus, the null hypothesis for the non-inferiority design was

$$H_0: p_T - p_C \geq 1.625\%$$

And the alternative hypothesis was

$$H_A: p_T - p_C < 1.625\%,$$

i.e., that the event rate with BV was at most 1.625% greater than the rate with heparin. Of course, they hoped that p_T would be less than p_C .

To test H_0 against H_A with a one-sided test at the 0.025 level and power of 0.80, the required sample size was 4,027 patients in each arm.

The ACUITY Trial

ACUITY enrolled 4,600 patients in each arm. The composite ischemic event rates were 7.7% in the BV arm and 7.3% in the heparin arm, a difference of 0.4% in favor of heparin.

Using standard methods, the 95% confidence interval for the difference in event rates was $0.4\% \pm 1.1\%$ or $(-0.7\%, 1.5\%)$. Thus, the one-sided 97.5% CI is $(p_T - p_C < 1.5\%)$. The NI criterion was met and the investigators rejected H_0 .

Expressed in terms of relative risk, the point estimate was 1.08 and the 95% CI was (0.93, 1.24)

The investigators concluded that “bivalirudin was associated with rates of ischemia ... that were similar to those of heparin.”

However, the FDA did not initially approve labeling for this use of bivalirudin on the grounds that the non-inferiority margin of 25% was unacceptably large. We will return to this issue.

Threats to Validity of NI Designs

Non-inferiority designs have some intrinsic limitations that make them more difficult to design and more vulnerable to problems than the superiority design. In the next few slides, we discuss the goals of NI designs and the assumptions required to ensure their validity.

We discuss the following issues:

- Implicit Goals of NI Designs

- Choosing the Non-inferiority Margin

- Assay Sensitivity

- Assay Constancy

- Efficacy Creep

Goals of NI Designs

The NI design has one explicit goal and one implicit goal. The explicit goal is to demonstrate that T is as effective, or nearly as effective, as best available therapy, C. The second goal is to demonstrate that T is better than P, that is, no treatment at all. Ordinarily, both must be true for T to be a therapeutic option.

In the setting of the ACUITY trial, the ideal would be a three-arm design with placebo, heparin, and BV. That would allow a direct comparison of T to C and of T to P. However, in the setting of acute coronary syndrome, a placebo arm would be unethical.

The non-inferiority trial provides a direct comparison of T to C, but it does not provide a direct comparison of T to P. Thus, one hopes to choose a non-inferiority margin that will provide assurance that T is better than P.

Choosing the Non-inferiority Margin, M

In the typical situation, previous research has shown that C is superior to P. A common practice is to combine all available evidence, perhaps through a meta-analysis, to estimate Δ , the treatment effect of C relative to P, and a confidence interval for Δ . We choose the NI margin based on the estimate of Δ .

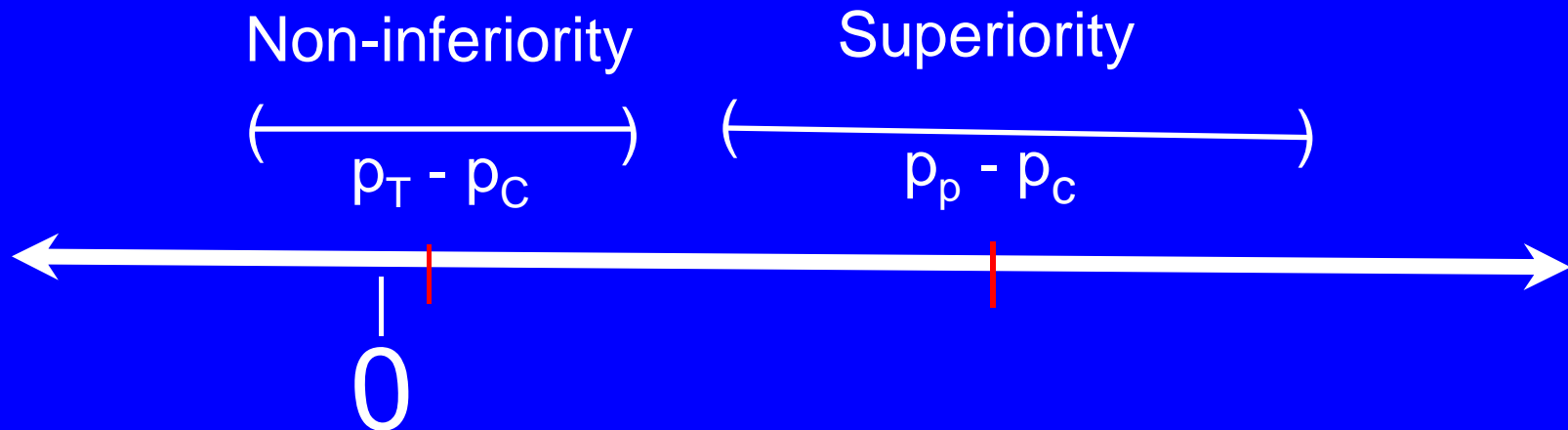
Since there is no single “correct” way to choose the NI margin, a host of methods have been proposed.

Set M equal to half the point estimate of Δ . The intent of this approach is to provide assurance that T provides at least half as much benefit as C.

From a statistical perspective, the margin should be, at the very least, no larger than the worst limit of 95% confidence interval (CI) of standard treatment effect relative to placebo, but it could be smaller so as to have assurance that the new treatment has greater than minimal efficacy.

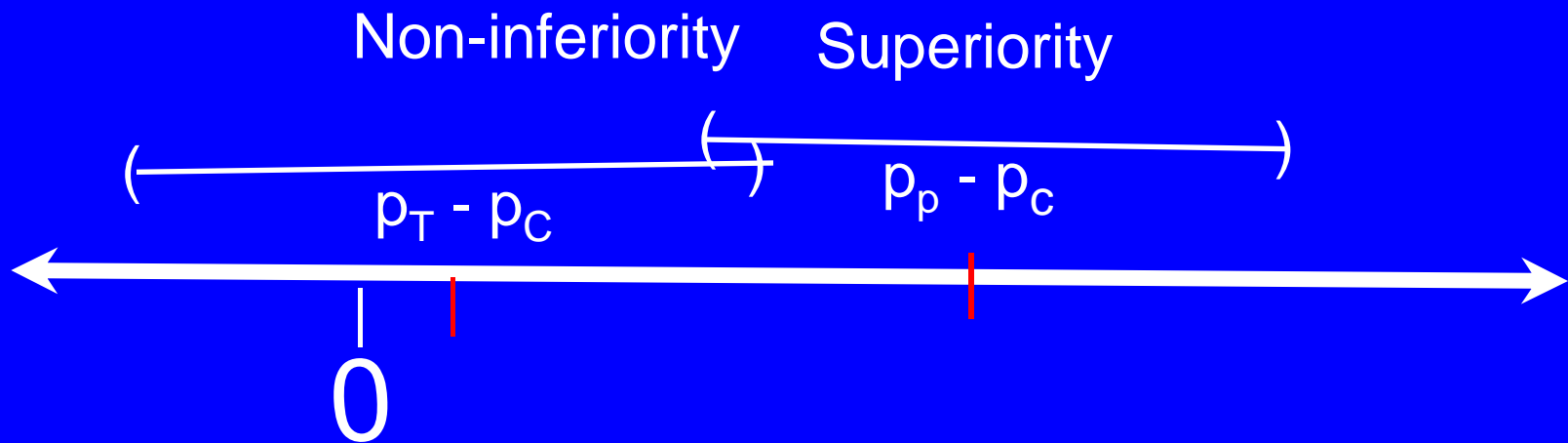
One proposal for selecting the margin is to take one-half of the magnitude of the worst limit of this CI—the so-called "50% rule" or "95-95 method." This conservative margin, however, often results in a high "false-negative" rate (type II error; i.e., low power to demonstrate non-inferiority)

Assuring T better than P



Provides some confidence
that $p_T < p_p$

Assuring T better than P



Less confidence that $p_T < p_p$

Choosing M

Clinical judgment must play the central role in the determination of the NI margin. Clinicians consider what difference in event rates would make the two treatments no longer “therapeutically equivalent”.

Because the requirement is so vague, pharmas meet with the FDA in advance of NI trials to seek agreement on the NI margin. Smaller NI margins imply larger sample sizes. For example, reducing the NI margin from 25% to 20% increases the sample size by about $(5/4)^2 = 1.56$. In the case of the ACUITY trial, no agreement was reached in advance and debate continues about whether the trial demonstrated non-inferiority.

Assay Sensitivity

A trial that demonstrates “equivalence” does not by itself demonstrate efficacy. Both treatments could be effective or both could be ineffective in the setting of the current trial.

The assumption that the comparator treatment is effective in the current trial is known as **assay sensitivity**. Many factors can contribute to assay insensitivity, e.g., 1) lower than expected event rates, 2) poor adherence, e) use of concomitant medications, and 4) spontaneous improvement of study patients.

A related idea, that the size of the treatment effect is the same in the current study and in the historical trials demonstrating the efficacy of C, is known as **assay constancy**. This is even more problematic because treatment of HIV evolves rapidly.

Properties of NI Trials

Ellenberg and Temple⁶ make two interesting observations about the NI design:

1. The incentive to reduce errors is reduced in NI trials. Since the goal of the trial is to show that the two treatment strategies have similar event rates, factors that reduce the differences between treatments, for example non-adherence or similar patterns of use of concomitant drugs, will increase the likelihood of success.
2. NI trials differ from superiority trials in that the interpretation of their results rests on external evidence that C is effective. In contrast, the interpretation of a superiority trial depends entirely on internal evidence of its design and results.

Efficacy Creep

Efficacy creep (sometimes called Biocreep) is the phenomenon that occurs when a slightly inferior treatment becomes the standard for the next generation of non-inferiority trials. If this happens repeatedly, one can end up with a standard that is no better than placebo.

To defend against efficacy creep, the active comparator should always be the “best” available therapy. However, the available evidence doesn’t always provide a clear choice of the best treatment.

Forum Question: Is it acceptable to use treatment other than best available therapy if BAT is too expensive for general use?

Summary

Because new treatments might represent advances despite increasing efficacy, NI trials are here to stay.

The International Conference on Harmonization guidelines say that “A suitable active comparator should be a widely used therapy whose efficacy has been clearly established and quantified in well-designed superiority trials and which can be expected to have similar efficacy in the contemplated trial”, and

“The margin in a non-inferiority trial is based on both statistical reasoning and clinical judgment, should reflect uncertainties in the evidence, and should be conservative.”

For a well-written summary of the issues, read the paper by D’Agostino et al¹⁰.

References

1. World Medical Association. *The Declaration of Helsinki*. <http://www.wma.net/e/policy/b3.htm>
2. The Clopidogrel in Unstable Angina to Prevent Recurrent Events Trials Investigators. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. *N Engl J Med* 2001;345:494-502.
3. Stone GW., McLaurin BT, David A. Cox DA, et al. Bivalirudin for Patients with Acute Coronary Syndromes. *N Engl J Med* 2006;355:2203-16.
4. Eikelboom JW, Anand SS, Malmberg K, Weitz JI, Ginsberg JS, Yusuf S. Unfractionated heparin and low molecular-weight heparin in acute coronary syndrome without ST elevation: a meta-analysis. *The Lancet*, 2000. 355:1936-42.
5. Snappin, SM Alternatives for discounting in the analysis of non-inferiority trials. *J Biopharm Stat* 2004;14:263-73.
6. Ellenberg SS, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1. Ethical and scientific issues. *Ann Intern Med* 2000;133:455-63.
7. Moliterno DF, Steven J Yakubov SJ, DiBattiste PM et al. Outcomes at 6 months for the direct comparison of tirofiban and abciximab during percutaneous coronary revascularization with stent placement: the TARGET₂ follow-up study. *Lancet*. 2002;360:355-360
8. Albers GW, Diener HC, Frison L, et al. Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation: a randomized trial. *JAMA*. 2005; 293: 690–698.
9. Sacco RL, Diener HC, Yusuf S, et al. Aspirin and extended-release dipyridamole versus clopidogrel for recurrent stroke. *N Engl J Med* 2008; 359:1238-1251.
10. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority Trials: Design concepts and issues – the encounters of academic consultants in statistics. *Stat Med* 2003;22:169-86.
11. Gao P, Ware JH. Assessing non-inferiority: A combination approach. *Statist in Med* 2008; 27:392-406.

The PEARLS Study: A Multinational Clinical Trial of HIV Treatment



Prospective
Evaluation of
Antiretrovirals in
Resource
Limited
Settings

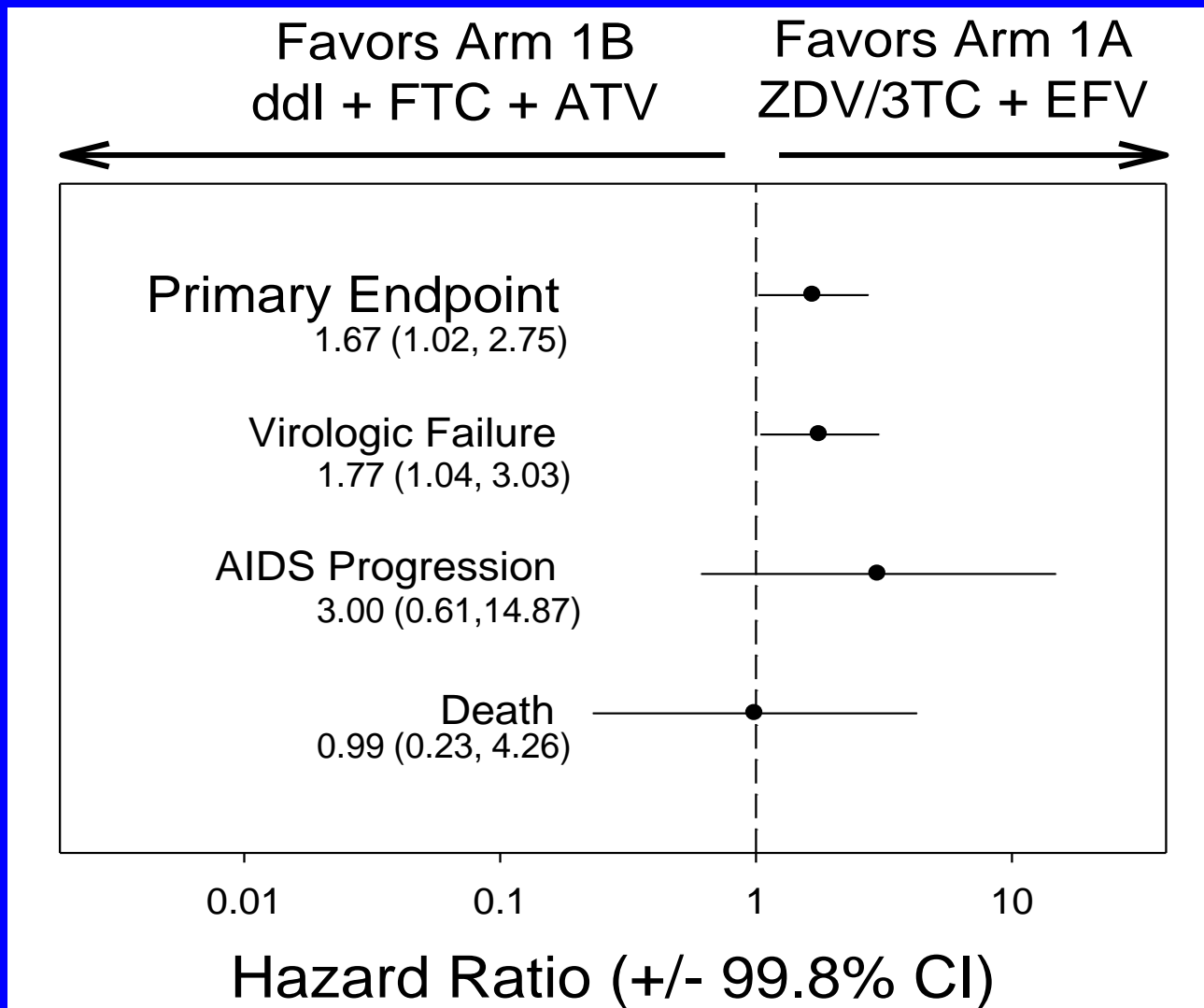
PEARLS Study Design

- Step 1 (initial regimen) 1:1:1 randomization
 - Arm 1A: ZDV/3TC BID + EFV QD
 - Arm 1B: ddl QD + FTC QD + ATV QD
 - Arm 1C: TDF/FTC QD + EFV QD
- Planned follow-up: the longer of 2.5 years or when at least 30% of participants have met the primary endpoint
- Want to show regimens 1B and 1C noninferior to standard 3-drug regimen, 1A:
- (95% Upper Conf. Bound of Risk Ratio < 1.35)

PEARLS Primary Endpoint

- Time to treatment failure defined as the time from randomization to first occurrence of any of the following:
 - Death: any cause
or
 - Disease progression - new or recurrent AIDS-defining OI or malignancy after 12 weeks of treatment
or
 - Virologic failure - plasma HIV-1 RNA \geq 1,000 copies/mL after 16 weeks of treatment

DSMB Findings (May 2008)



Knowledge Gained

- Comparison of efficacy and safety of 3 different ARV regimens
 - Each has potential for use in resource-limited setting
 - Expected availability after study completion
- Better understanding of how ARVs interface with unique features of each community:
 - Endemic co-infections
 - HIV strains (subtypes)
 - Nutrition
 - Human genetics
 - Human behavior
 - Traditional medications
- Dissemination of knowledge to the communities

Appendix

Supplementary Slides

What is the Problem?

We¹¹ have argued that these difficulties arise because investigators fail to recognize that NI trials have two separate objectives, namely:

O₁: The trial must demonstrate that T is “therapeutically equivalent” to C. That is, we must show NI using a criterion that meets the standard for an acceptably largest difference in efficacy.

O₂: The trial must also demonstrate that T is superior to P, with appropriate consideration of the uncertainty associated with the historical evidence.

These two objectives can be achieved only by designing the trial to test two separate hypotheses.

Testing for Non-Inferiority

To address O_1 , that is, non-inferiority, we determine a non-inferiority margin, M , which, if met will allow us to conclude that T is therapeutically equivalent to C . The margin should be chosen based on clinical considerations, not historical data regarding the effect of C relative to P .

Given the margin, M , we test the null hypothesis

$$H_{10}: p_T - p_C \geq M$$

Against the alternative hypothesis

$$H_{1A}: p_T - p_C < M$$

using standard methods for NI trials.

Testing for Superiority to Placebo

A direct comparison of T to P is not available in a non-inferiority trial. If, however, we believe that the effect of C relative to P is identical in the contemporary and historical contexts (assay constancy), we can represent $p_T - p_P$ in terms of parameters from the placebo-controlled trial of C and from the non-inferiority trial, namely

$$p_T - p_P = (p_T - p_C) - (p_C - p_P)$$

The assumption of assay constancy is a source of considerable concern at the FDA, so we will return to this point.

Testing for Superiority to Placebo

Assuming assay constancy, we test the null hypothesis:

$$H_{20}: \Delta_{TP} (= \Delta_{TC} + \Delta_{CP}) = 0$$

against the alternative $H_A: \Delta_{TP} < 0$ with the test statistic

$$\frac{\hat{\Delta}_{TC} + \hat{\Delta}_{CP}}{\sqrt{\hat{V}_{TC} + \hat{V}_{CP}}}$$

If we assume approximate normality of the estimates, this can be done with standard methods.

Assay Constancy

To address concerns about assay constancy, we consider a variant of this method based on the **discounted synthetic estimate**. Suppose we assume that the effect of C relative to P in the contemporary setting is only a fraction $(1 - \lambda)$ of the effect in the historical setting. Then

$$\Delta_{TP,\lambda} = \Delta_{TC} + (1 - \lambda)\Delta_{CP} = 0$$

We then test H_{20} : $\Delta_{TP,\lambda} = 0$ with the test statistic

$$\frac{\hat{\Delta}_{TC} + (1 - \lambda)\hat{\Delta}_{CP}}{\sqrt{\hat{V}_{TC} + (1 - \lambda)^2 \hat{V}_{CP}}}$$

Design and Analysis

Since the goal of the NI design is to reject both H_{01} , the hypothesis of inferiority, and H_{02} , the hypothesis of no difference relative to placebo, we perform sample size calculations for each test and choose N equal to the larger of the two sample sizes.

The analysis does not require two unrelated tests. Since the results from the historical trials are known at the time the NI trial is conducted, the two hypothesis tests can be reduced to two tests involving the estimated treatment effect of T relative to C in the NI trial. The two tests can be expressed as inequalities involving

$$\hat{\Delta}_{TC}$$

If the estimate satisfies the more stringent of the two inequalities, we reject both H_{01} and H_{02} and achieve both objectives.

Implications

A particular clinical trial might achieve only one of the two objectives.

Suppose that we demonstrate that T is superior to placebo (O_2) but fail to show that T is therapeutically equivalent to C (O_1). The drug might be of use in the treatment of patients for whom C is contraindicated.

In contrast, if we achieve O_1 but not O_2 , we can conclude that T is therapeutically equivalent to C but may not be superior to P. This is likely to arise when the evidence supporting C is weak. This is a dilemma if a placebo-controlled trial is considered unethical.

The CURE Trial

The CURE Trial² compared clopidogrel plus aspirin to placebo plus aspirin for the treatment of patients with acute coronary syndrome without ST-segment elevation.

The investigators believed that clopidogrel, a thienopyridine derivative, would reduce the incidence of ischemic events by inhibiting platelet aggregation.

To test this hypothesis, they specified a **null hypothesis**, that clopidogrel and placebo treatment would produce the same one-year incidence of death, MI, or stroke. The **alternative hypothesis** (H_A) was that clopidogrel would reduce the one-year incidence of ischemic events.

The hope was that the data would lead to rejection of the null hypothesis and the conclusion that clopidogrel is superior.

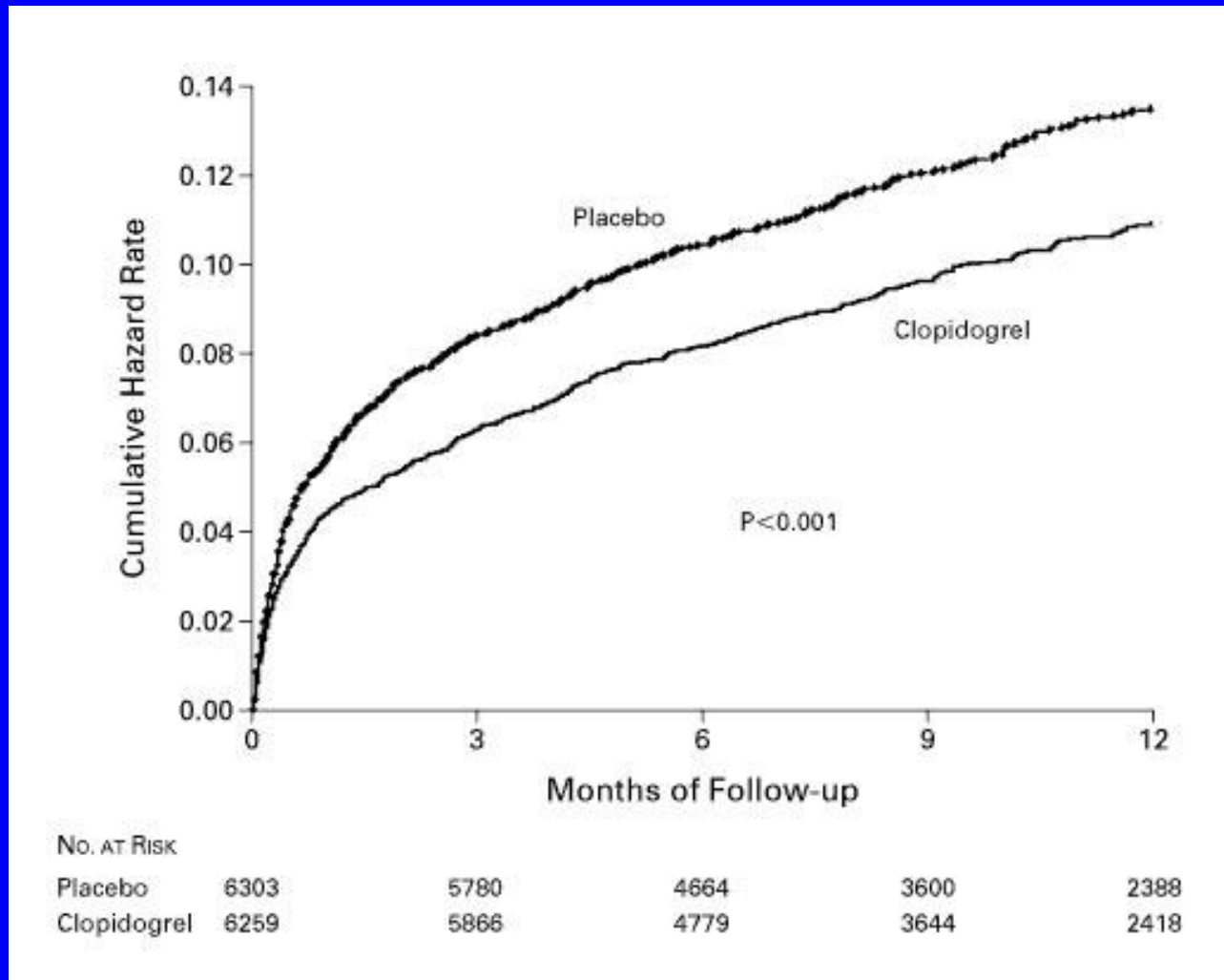
The CURE Trial

Assuming an event rate of 10% in the placebo group and a 17% reduction in the event rate in the clopidogrel group (to 8.3%), the investigators determined that 12,500 patients would be needed to test the null hypothesis with Type 1 error of 0.045 and power of 0.90.

The observed results were one-year event rates of 9.3% in the clopidogrel arm and 11.4% in the placebo arm ($P < 0.001$).

The investigators rejected the null hypothesis and concluded that clopidogrel is **superior** to placebo for the prevention of death, MI, or stroke for patients with acute coronary syndrome without ST-segment elevation.

Cumulative Hazard Rates for the First Primary Outcome (Death from Cardiovascular Causes, Nonfatal Myocardial Infarction, or Stroke) during 12 Months of Treatment



The Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial Investigators. N Engl J Med 2001;345:494-502

Results from Meta-analysis of Trials Comparing Heparin to Aspirin

Eikelboom et al⁴ performed a meta-analysis to estimate the relative efficacy of heparin regimens to aspirin in patients ACS. They identified 12 trials involving a total of 17,157 patients.

Based on the data from these trials, they estimated the relative risk of death or MI during short-term treatment (up to 7 days) as 0.53 with a 95% confidence interval of (0.38, 0.73).

Performing the calculations on the natural logarithmic scale as is customary for relative risk, half of the estimated benefit is 0.73. To preserve this benefit, we must have $0.73 \cdot M < 1$ or $M < 1.37$.

So a non-inferiority margin of 1.25 seems reasonable by this criterion.

Choosing the Non-inferiority Margin, M

However, we know that the if the estimate of Δ is subject to uncertainty. Perhaps the true effect was actually smaller than Δ

FDA statisticians proposed that M be set equal to half the benefit based on the upper limit of the confidence interval for Δ from a meta-analysis of previous studies. This is known as the **95-95 method**⁵ and is intended to insure that 50% of the smallest possible benefit is preserved. Half the upper limit of 0.73 is 0.85, so the non-inferiority margin by this method is $1/0.85 = 1.18$. By this criterion, the non-inferiority margin for ACUITY should be 1.18.

This method is conservative in that it provides strong assurance that T is superior to P if the NI trial is successful. It also requires very large, sometimes impossible, sample sizes.

The substantive point is that there is no objective, agreed method for choosing M. This leads to controversy and difficulties in the regulatory setting.

Choosing the NI Margin: Other Examples

In the TARGET⁷ trial, two glycoprotein IIa/IIIb inhibitors, tirofiban and abciximab, were studied to establish the non-inferiority of tirofiban in prevention of cardiovascular events in patients undergoing percutaneous coronary revascularization (stenting).

The primary endpoint was death, MI, or urgent revascularization in 30 days.

The non-inferiority margin for the hazard ratio was chosen as 1.47, half the effect of abciximab in the EPISTENT trial. The estimated hazard ratio and its 95% confidence interval were 1.26 (1.01, 1.57). The investigators concluded that they demonstrated T to be inferior to A.

The trial was later judged to be poorly designed because an agent with a hazard ratio as large as 1.47 would not have been considered to be therapeutically equivalent to abciximab.

Choosing the NI Margin: Examples

In the P_{Ro}Fess Trial⁹, the investigators sought to demonstrate the non-inferiority of aspirin plus dipyridamole relative to an active control, clopidogrel, for the prevention of recurrent stroke.

Following the 95-95 approach suggested by FDA statisticians, they chose a NI margin equal to half the lower limit of the confidence interval from placebo-controlled trials of clopidogrel. This gave a NI margin for the hazard ratio of 1.075.

To achieve a manageable sample size, the investigators chose an alternative hypothesis that T would reduce the event rate by 6.5%, that is, that T would actually be superior to C.

When the trial was carried out, the observed event rates were almost identical [9.0% (A+D) and 8.8% (C), RR = 1.01 (0.92, 1.11)], but the data did not satisfy the pre-specified NI criterion.

Choosing the NI Margin: Examples

Similarly, in the SPORTIF trials⁸, ximelegatran was compared to warfarin for stroke prevention in patients with atrial fibrillation patients. Again based on the historical evidence, the sponsor chose an absolute non-inferiority margin of 2%.

The event rates in the warfarin group (control) were 2.3% (Sportif III) and 1.2% (Sportif V). Because of the low event rates in the control arm, this resulted in a margin that allowed the conclusion of non-inferiority even with a doubling of the event rate in the X arm.

The common theme in these trials was that a non-inferiority margin based on the effect of the comparator drug was not consistent with the clinical standard for therapeutic equivalence.

In response to these and other trials, the FDA began to take a more conservative stance about the choice of the NI margin. FDA statisticians advocate routine use of the 95-95, or two confidence-interval, method.

Background

Early on, most clinical trials were placebo-controlled. The experimental treatment, T, was compared against a placebo, P, possibly in the context of other treatment. The goal of the trial was to show that T is **superior** to P. However, placebo-controlled trials are not appropriate when effective treatments have been identified. The ethical issue was illustrated dramatically by the infamous Tuskegee Study of patients with syphilis.

The Declaration of Helsinki¹ states that “The benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current ...therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where no proven ... therapeutic method exists.”

This precludes the use of a placebo control if that would require withholding a proven therapy.