FDA Experience and Perspective on Non-Inferiority Trials

Robert J. Temple, M.D.
Associate Director for Medical Policy
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

FDA Workshop on CAP
January 18, 2008

Introduction

You've heard a lot about NI trials already, so I'll try not to repeat. The critical issues are, by now known to you:

1. In most cases NI trials pose inferential problems, but you use them when you have no choice, i.e., when you simply cannot leave patients untreated (placebo-treated) and must use active treatment as the control. But the need to use an active control does not always mean this design will be a valid test of effectiveness.

Introduction (cont)

2. The NI study seeks to show that the new drug is not inferior to the standard by too large an amount; that amount is called the non-inferiority margin, M or delta. The NI margin has two determinants

- Inferiority must not be greater than the whole effect of the control (because then you've lost the whole effect). So you must know the effect of the control in the new study. The whole effect of the control is called M_1 the largest possible NI margin.

- Inferiority must not be clinically unacceptable. This is a clinical, not statistical, judgment. The largest clinically acceptable difference is called M_2. It must be no longer than M_1.

Introduction (cont)

3. The critical problem in the NI trial is "assay sensitivity" (AS). Is this a trial that could have detected the difference of interest if there were such a difference? To do that, the active control must have had an effect in this study of at least M_1. If it didn't, showing inferiority of the test drug less than M_1 (i.e., non-inferiority) will not tell you that the test drug has any effect and will be meaningless with respect to effectiveness.

Introduction (cont)

4. You don't actually measure the effect of the control, or assure assay sensitivity, in the NI study. You have to assume the size of the effect of the active control, based on past experience. And if you are wrong, and the active control did not have such an effect in this study, you could conclude that an ineffective treatment works.

This problem has long been recognized by some trialists and by FDA, and it creates uncertainty about the meaning of an NI trial.

Citation of Expert Opinion

In serious but less critical medical situations, one can justify a comparison between new drug and standard, even if a placebo group seems out of the question. But such a trial is convincing only when the new remedy is superior to standard treatment. If it is inferior, or even indistinguishable from a standard remedy, the results are not readily interpretable. In the absence of placebo controls, one does not know if the "inferior" new medicine has any efficacy at all, and

(continued)

"equivalent" performance may reflect simply a patient population that cannot distinguish between two active treatments that differ considerably from each other, or between active drug and placebo. Certain clinical conditions, such as serious depressive states, are notoriously difficult to evaluate because of the delay in drug effects and the high rate of spontaneous improvement, and even known remedies are not readily distinguished from placebo in controlled trials. How much solace can one derive from a trial that shows no difference between a new putative antidepressant and a standard tricyclic?

Lasagna, L: Eur J Clin Pharm
15:373-374, 1979

Problems of Active Controlled Trials

As early as 1982, FDA regulations recognized the fundamental problem of the trial seeking to show similarity, namely the necessary assumption of ASSAY SENSITIVITY, i.e. an assumption that the trial could have detected a difference of specified size between two treatments if there were one. The regulation said

"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 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."

Problems of Active Control Trials

So, for more than 20 years, the major problem with the equivalence or non-inferiority design has been recognized and the general description of the potential solution known: you have to analyze the past performance of the active control to know whether it can be assumed to have an effect of defined size in the new study.

This critical assumption gives non-inferiority studies an unsettling similarity to historically controlled studies. In those you must be able to say, from past observations, what would happen to an untreated group of patients like those in the current study. In the non-inferiority study you need to say what the effect of the control drug in the new study would have been compared to a placebo.

That can be very difficult

Assay Sensitivity and Choice of NI Margin

NI trials once were called "equivalence" trials. Often these consisted of comparing two drugs, showing "no significant difference" and declaring "equivalence/victory."

But you can only really show equivalence by being superior and no significant difference can mean too small a study and many other things. We now ask that the difference (degree of inferiority) of the test drug to the control (C-T) be smaller than some margin (M).

$$\text{i.e., } C-T < M,$$

where M can be no greater than the entire effect of C in this study. If the difference, C-T, is < M, then T has some (> 0) effect.

So it's really a "not too much inferiority" study.

The analytic methods are familiar from standard placebo-controlled trials.

The Logic of the Non-Inferiority Trial

In a placebo-controlled trial, the null hypothesis is that the test drug T is ≤ O.

$H_o: T \leq O$

$H_a: T > O$

This is established by showing that the 97½% lower bound of the CI for T-placebo is >O.

The Logic of the Non-Inferiority Trial

In the non-inferiority study, the null hypothesis is that the degree of inferiority of the new drug (T) to the control (C), C-T, is greater than the margin M

$$H_o: C\text{-}T \geq M \text{ (T is more inferior than M)}$$

$$H_a: C\text{-}T < M \text{ (T is less inferior than M)}$$

For the study to show an effect of T, M can be no larger than the whole effect of C in that study. Again you compare the 97½% CI upper bound of C-T with M.

M is Crucial

Everything depends on the validity of M, i.e., that you are sure that the effect of C in the new study is at least M.

M thus needs to be chosen conservatively. If, e.g., you say M=10, then if C-T (95% CI upper bound) is < 10, T has an effect. But if in the study the effect of C is only 5, T will NOT have had an effect.

IT WILL ONLY LOOK LIKE IT DOES

You need to be very sure of the margin

This leads to conservative choices and large sample sizes.

Problems of Non-Inferiority Studies

If the logic of an NI trial is OK, what's the problem: There are 3:

1. The assumption of Assay Sensitivity

   There is a critical assumption: that the trial could have detected a difference (or a difference of defined size), had there been one. This property, called Assay Sensitivity, in turn depends on the assumption that the control drug would have had an effect of at least some specified size <u>in this study</u> (compared to placebo) had there been a placebo group. But the effect of the control drug is not measured (there is no placebo group) and the assumption cannot be supported in many situations.

   N.B. This is not a matter of power. Power tells you what difference you could have detected. But if the difference you wanted to rule out is 5 (the margin M that you believe the control drug had in the study) and you in fact rule out a difference of 5 or more, that has no meaning if the effect of the control was actually only 2 (or zero) in this study. That study lacked Assay Sensitivity; it could not have detected a difference between the treatments that would have shown the new drug to have had no effect.

Fundamental Problems

2. Retaining more Than "Any" Effect
   The whole logic of the trial depends on showing that the difference between treatments (C-T) is less than some margin M_1, where M_1 is the whole effect of the control. That margin cannot be > the effect of the control drug. But the margin also must not be greater than a clinically critical difference M_2, where $M_2 \leq M_1$. After all, you're doing an active control trial because you don't want to leave people untreated. You also don't want them "barely treated." M_2 has to be chosen to reflect the clinical value of the drug. This can lead to very large sample sizes.

3. "Sloppiness Obscures Differences."
   The need to show a lack of difference (as opposed to some difference) can lead to lack of incentive to study excellence:

Assay Sensitivity

A property of a clinical trial: the ability to distinguish active from inactive drugs, or, in a specific case, the ability to show a difference of a specified size M between treatments where M is the effect of C that is presumed present in the new study. To do this, the control must have an effect at least M larger than no treatment. If the trial did not have assay sensitivity, then even if C-T < M, you have learned nothing about the effect of T.

If you don't know whether the trial had assay sensitivity, finding no difference between C and T means either that, in that trial:

Both drugs were effective
Neither drug was effective

Determining Assay Sensitivity

To conclude a trial had assay sensitivity, you need a combination of 1) historical information, 2) assurance of similarity of the new trial to historical trials, and 3) information about the quality of the new trial.

1. Historical evidence of sensitivity to drug effects (HESDE)

   A historically based conclusion that appropriately designed, sized, and conducted trials in a particular disease, with a specific active drug (or group of related drugs) reliably show an effect of at least some defined size on a particular endpoint. Usually established by showing that appropriately sized (powered) and well-conducted trials in a specified population regularly distinguish the active drug(s) from placebo for particular endpoints

   Sensitivity to drug effects is an abstract conclusion about well-designed trials of a drug in a particular disease. Assay Sensitivity is a conclusion about a particular trial

Determining Assay Sensitivity

1. HESDE

For most symptomatic treatments, history clearly does not suggest a new trial will have assay sensitivity; i.e., many well-designed studies fail to show effects

Anxiety	CHF symptoms
Depression	Angina
Insomnia	GERD Symptoms
Allergic rhinitis	Irritable bowel syndrome
Asthma prophylaxis	Pain

For some outcomes studies, results are also inconsistent, notably survival post-MI with beta blockers or aspirin

Could it be sample size? Maybe, but in these cases it looks as if some trials are different from others; i.e., there is a treatment by study interaction.

Cases Where As Is Pretty Certain

Unlike many symptomatic conditions, there are situations in which treatment responses are large, plainly different from placebo.

- Heparin in deep vein thrombosis
- Strep throat, UTI's
- Treatment of acute leukemia, testicular Ca
- Beta agonists in broncospasm

There are other cases where analysis showed very consistent results across studies

- Steroid asthma prophylaxis
- Thrombolytics in AMI

Determining Assay Sensitivity

2. Similarity of Current Trial to Past – the Constancy Assumption

Conclusion of HESDE applies only to trials of a particular design (patient population, selection criteria, endpoints, dose, use of washout periods and, particularly important, background therapy) . Changes in these can alter the effect size of the active control and, therefore, the appropriate margin, or completely undermine assay sensitivity

For example:

Effect on mortality of post-infarction beta blocker treatment could be altered by new medications (lipid lowering, anti-platelet drugs) or procedures (CABG, angioplasty)

Effect of ACEI on CHF could be altered by routine use of beta-blockers or aldosterone antagonists

Effect of a thrombolytic could depend on how many hours after onset of AMI treatment was started

Determining Assay Sensitivity

3. Study Quality

If sensitivity to drug effects exists for a therapeutic class, assay sensitivity in a particular study can still be undermined by a variety of study conduct factors that "bias toward the null," i.e., obscure true differences between treatments and cause the historical experience to represent an overestimate of the effect of the control

These factors include:

Determining Assay Sensitivity

3. Study Quality (cont.)

   - Poor compliance
   - Non-protocol crossovers
   - Spontaneous improvement in the population
   - A poorly responsive population
   - Use of concomitant medication that reduces potential response
   - Poor diagnostic criteria (patients lack the disease)
   - Inappropriate (insensitive) measures of drug effect
   - Poor quality of measurements
   - Mixing up the treatments

Overall there is a lower incentive to high quality in trials seeking to show no difference between treatments. History could therefore overestimate the effect of the control in the new trial

Determining Assay Sensitivity

3. Study Quality (cont.)

These factors, in general, have only small (or no) effects on variance (width of CI) but can reduce or obliterate C-T differences, leading to false conclusion of non-inferiority

Note: Some analytic approaches that are "conservative" in a difference-showing trial are not in a non-inferiority trial; for example, an intent-to-treat approach reduces C-T and is not conservative

M_2, the Clinical Margin

M_1 is the largest possible non-inferiority margin because it represents the entire effect of the control in the study. You need to rule out inferiority of T by $>M_1$ to be sure T has any effect at all. But if the effect is of value, assuring retention of any of the control effect may not be adequate. It is therefore common to choose M_2 as the non-inferiority margin, where M_2 represents the smallest effect (often thought of as a fraction of M_1) that must be preserved. Note that you cannot assure true equivalence or no inferiority at all except by having T be superior to C

Confusion of M1 and M2

There has been a tendency to consider M_1 and M_2 separately or more specifically to consider M_2 without reference to M_1. That is all right if $M_1 >> M_2$ (e.g., many antibiotic treatments, treatment of acute leukemia) where the effect is so large that the only issue really is comparative effectiveness, but not if M_2 is almost equal to M_1 (or larger). In the past it was common in cancer trials to declare equivalence if survival inferiority of 20% was excluded. But the control agent in many studies did not have a known effect as large as 20% more than no treatment (that's a 2 month survival advantage if the control is 10 months) so that successfully excluding a more than 20% difference could represent loss of all effect or even harm. In many cases this approach was used even if no survival effect of the control was documented

There is a certain logic to that approach regarding clinical value, but it cannot show <u>effectiveness</u>

Confusion of M_1/M_2 (cont)

The oncology experience has been replicated in ID. It seems pretty clear that the "clinically insignificant" 10-15% differences used as margins in otitis, sinusitis, and acute exacerbations of chronic bronchitis, while perhaps truly insignificant, were larger than the usual effect of the control agent. These margins could not show drug effectiveness.

So it is absolutely critical to rigorously define M_1 before considering what portion of it must be retained.

Choosing the Margin

1. Similarity in some ways to historical controls

The need to assume assay sensitivity and control drug effect size gives all non-inferiority studies the unsettling element of a historical control, a kind of control with well-recognized problems. (It is just as critical to believe you understand what the present effect of the control drug is, compared to a placebo, when you define M based on past experience, as it is to believe you know what would happen to a current untreated group of patients based on a past experience, so you can compare new treatment with old untreated).

Choosing the Margin (M_1)

2. Historical results may be variable

Establishing HESDE demands that there be a complete look at past placebo-controlled trials. If results are consistent, there's generally no problem

But if results vary widely, the choice of margin is difficult. It obviously cannot be based on the most favorable single result because that would overestimate the control drug effect in the current trial, but it also cannot be based on the point estimate of a meta-analysis because may trials have poorer results. In this case, even the lower bound of a 95% CI is a problem because this could be greater than the actual result in some studies. And results can vary widely, as experience with GP IIb/IIIa inhibitors shows

Choice of Margin (M_1) (cont.) Variability Can Be Great

Example: IIb/IIIA antagonists (Abciximab, eptifibatide, tirofiban) for use after percutaneous intervention (Kong/Califf)

All: (48-96 hrs), 43% reduction in death, NFMI, but individual results varied from

>50% (Impact III, CAPTURE, several others)

<25% (EPILOG, RAPPORT, RESTORE, PRISM+)

Given values of <25% in good sized trials, is 43% (with variance) a good representation? It would be much easier to rule out loss of a 43% effect than a 25% effect, but there's a good chance, for a given study, that the effect is < 43%

Choice of Margin (M_1) (cont.) Variability Can Be Great

Surely would prefer to use data on single drugs rather than pooled "pharmacologically similar" drugs, as drugs within a "class" can differ. But this will greatly widen the confidence interval and 95% CI lower bound.

Variability (cont.)

IIB/IIIA inhibitors are even more variable for ACS, with early reduction of 29% in death and AMI but abciximab (which tends to be best after PCI) in GUSTO IV showed no effect at all, a major surprise

Every trial differs with respect to precise definition of patients, kind of heparinization, indicators for initial or further intervention

If one accepted active control, might want to (1) choose drug that is numerically best for the control agent, and (2) choose margin at low end, not average

Choice of Margin

Response to Limitations of Data

The variability in trials, modest event rates, and generally cautious choices of M_1 pose significant problems. Many findings drive M_1 toward lower value.

- If range of effect sizes, and one study is planned, plainly need to "go low"
  - It's very hard to choose a value for M that is larger than the lower bound seen in an actual study
  - Even one failed study is a major problem (yet there are failed studies for beta blockers, aspirin, etc)

Choice of Margin
Changes in Response Over Time

3. Even if past trials regularly show superiority to placebo, effect size could differ from the past. The " constancy assumption may be shaky

- Results could depend on how soon treatment started (hours, thrombolytics; days, beta blockers
- Therapy will often have changed (effects of ACEIs in CHF were without beta blockers, spironolactone, statins)
- Trials may be done in a different region (is NYHA CHF score the same everywhere). We know relatively little about consistency across regions.

The Special Case of $M_2 << M_1$

The difficulties of setting the margin are clearly less if $M_2 << M_1$. In that case you don't need to be very precise about effect size. If, e.g., cure rates are 80% at one week in UTI with treatment and 20% without, and M_2 is 10-15%, you don't need to worry much about the absolute cure rate of the control in the study.

It's where effect size (M_1) is in doubt, uncertain, absent in some studies, that problems emerge.

In the present case a firm conclusion for a defined category of CAP that the effect of treatment is, say, 30%, should make life easy for an M_2 of 10%.