U.S. Food and Drug Administration

Notice: Archived Document

The content in this document is provided on the FDA's website for reference purposes only. It was current when produced, but is no longer maintained and may be outdated.

FDA's Clinical Investigator Course

Cosponsored by

FDA's Office of Critical Path Programs (OCPP)
and
The Clinical Trials Transformation Initiative (CTTI)

U.S. Department of Health and Human Services
Food and Drug Administration

Design of Clinical Trials

Robert Temple, M.D.
Food and Drug Administration
November 8, 2011

Outline

1. Adequate and well-controlled studies and the 1962 FD&C Act

2. Non-inferiority studies

3. Enrichment designs

4. Dose response (probably no time)

The Effectiveness Requirement

Until 1962 drugs had to be shown “safe” to be marketed, but there was no requirement to show effectiveness. There was talk about effectiveness (how can a drug be safe if it provides no benefit) but no requirement and few studies we would recognize as useful.

Then it all changed with the 1962 amendments to the FD and C Act.

The Effectiveness Requirement

An NDA can be rejected if:

There is a lack of substantial evidence that the drug will have the effect it purports or is represented to have under proposed labeled conditions of use (this is what an applicant must show)

The Law then goes on to describe what substantial evidence is. It is evidence consisting of adequate and well-controlled investigations, including clinical investigations…on the basis of which it could be concluded that the drug will have the effect it is represented to have under the conditions of use proposed in labeling (this is how the applicant must show effectiveness)

The Effectiveness Requirement

It was the only new requirement for approval in 1962

It was not the effectiveness requirement that was radical. I believe we might have imposed that by regulation, as “safe for intended use” could alone imply a risk/benefit analysis, i.e., need for evidence of benefit; It was the need for adequate and well-controlled studies that changed <u>everything</u>, all of medical science, really

- These are the <u>only</u> basis for approval
- Note the plural. Agency interpreted this as requiring <u>more than one</u> controlled trial (modified by FDAMA 1997 to allow one study in some cases)
- No relative efficacy (unless inferior effectiveness leads to lack of safety)
- Effect must be clinically meaningful (added by Federal court)

The Effectiveness Requirement (cont.)

It was really an amazing stroke

- In those days (not any more), laws tended to be general, leaving details to the agencies with expertise. That philosophy might have led to a substantial evidence requirement, not further defined

- For Congress to go further and say what the only kind of acceptable study could be was remarkable

- Actually a very clever trade-off. "Substantial," legally, is a low standard (between a scintilla and a preponderance)

  But adding a need for two A&WC studies turns a low standard into quite a high one [especially with the $p<0.05$ (two-sided) that emerged]

The Effectiveness Requirement (cont.)

In 1962, of course, and really until 1970 or so, we at FDA had only a poor idea of what a well-controlled study was, and things we take for granted now were not at all known. But we have learned and learned, about the importance of interim looks, maintaining blinding, multiplicity, the importance of good dose-response, the difficulties of active control trials, and much, much more. I will touch on some of these experiences.

Adequate and Well-Controlled Studies

314.126 Adequate and Well-Controlled Studies

Only basis for approval

Apart from design and analysis (A and WC) must show effectiveness convincing to experts, ordinarily a statistically significant effect on a meaningful endpoint.

Adequate and Well-Controlled Studies

Directed at three main goals:

1. Need a valid control group because the course of a disease is variable; the state of the disease can change spontaneously and is subject to many influences. The control group is a group very similar to the test group and is treated the same as people getting the test drug, except for getting the drug. It lets you tell drug effect from other influences, such as spontaneous change, placebo effect, biased observation.

(If course was predictable, you would just intervene and observe.)

Adequate and Well-Controlled Studies

Main Goals

2. Need to minimize bias, a “tilt” favoring one treatment group, a directed (non-random) difference in how test and control group are selected, treated, observed or analyzed

3. Sufficient detail to know how the study was done and what results were

These goals are set forth in detail in regulations at 21 CFR 314.126.

Adequate and Well-Controlled Studies (Cont'd)

Reports of adequate and well-controlled investigations provide the primary basis for determining whether there is "substantial evidence" to support the claims of effectiveness for new drugs and antibiotics. Therefore, the study report should provide sufficient details of study design, conduct, and analysis to allow critical evaluation and a determination of whether the characteristics of an adequate and well-controlled study are present.

Adequate and Well-Controlled Studies (Cont'd)

(B) An adequate and well-controlled study has the following characteristics:

(1) There is a clear statement of the objectives of the investigation. In addition, the protocol should contain a description of the proposed methods of analysis, and the study report should contain a description of the methods of analysis ultimately used. If the protocol does not contain a description of the proposed methods of analysis, the study report should describe how the methods used were selected.

Adequate and Well-Controlled Studies (Cont'd)

(2) The study uses a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect. The protocol for the study and report of results should describe the study design precisely; for example, duration of treatment periods, whether the treatments are parallel, sequential, or crossover, and whether the sample size is predetermined or based upon some interim analysis. Generally, the following types of control are recognized:

Kinds of Controls

Placebo control

No treatment concurrent control

Dose-response control

Active Control

Historical Control

There is no “hierarchy;” all types can be, and in any given year are, used as the basis for approval of a drug. But not every design is usable in every situation.

Difference-Showing vs. Equivalence/NI

Difference showing trials

- Placebo control
- No treatment
- Dose-response
- Some active control
- Most historical control

Non-Inferiority-showing trials

- Most active control
- Some historical control

Adequate and Well-Controlled Studies (Cont'd)

(I) Placebo Concurrent Control. The test drug is compared with an inactive preparation designed to resemble the test drug as far as possible. A placebo-controlled study may include additional treatment groups, such as an active treatment group or more than one dose of the test drug, and usually includes randomization and blinding of patients or investigators, usually both.

Ethics
Difference-showing
Blinded, randomized
No external data needed (assay sensitivity)

Baseline placebo
Add-on studies
Randomized withdrawal

Adequate and Well-Controlled Studies (Cont'd)

(II) Dose-Comparison Concurrent Control. At least two doses of the drug are compared. A dose-comparison study may include additional treatment groups, such as placebo control or active control. Dose-comparison trials usually include randomization and blinding of patients or investigators, or both.

Effectiveness vs. D/R

Dose-Response

D/R study one kind of controlled trial

Growing recognition that it is important to choose a reasonable dose - ICH guideline 1993.

- Historical error: diuretics
- Effective dose 1/8-1/4 dose used
- Hypokalemia, probably decreased benefit of treatment
- Disparity between stroke effect (40%) and cardiac effect (15%) until low-dose used (SHEP)

Goal: Define D/R curve for benefits and risks

Dose-Response Studies

Until early 1980's, most trials with more than one dose titrated the dose, generally to some endpoint. This meant:

1. The group on any given dose was not chosen randomly

2. Time and dose were confounded; secular trend would look like response to dose. Particularly useless for safety

In 1980's, FDA promoted the randomized, parallel, fixed dose, dose-response study, identified as the standard in ICH E4 guidance. Note, D/R studies can serve two purposes:

1. Show effectiveness
2. Show D/R

Adequate and Well-Controlled Studies (Cont'd)

(III) No Treatment Concurrent Control. Where objective measurements of effectiveness are available and placebo effect is negligible, the test drug is compared with no treatment. No treatment concurrent control trials usually include randomization.

Like placebo but unblinded

GUSTO, GISSI, cancer trials

Objective endpoints (ART, LRC)
Unblinded additional Rx

But be careful about what is "objective"

Recent concern regarding the RECORD study of Avandia, referrals of cases for adjudication.

Adequate and Well-Controlled Studies (Cont'd)

(IV) Active Treatment Concurrent Control. The test drug is compared with known effective therapy; for example, where the condition treated is such that administration of placebo or no treatment would be contrary to the interest of the patient. An active treatment study may include additional treatment groups, however, such as a placebo control or a dose-comparison control. Active treatment trials usually include randomization and blinding of patients or investigators, or both. 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.

Equivalence/Non-Inferiority Trials

A major regulatory, ethical, international problem

Fundamental distinction between trials intended to show a difference and trials intended to show similarity; latter pose major problems of interpretation.

Desire to use equivalence/NI is understandable: seems sensible to compare new and old effective therapy, see no difference and declare victory. Avoids exposure to ineffective treatment.

I will return to this shortly.

Adequate and Well-Controlled Studies (Cont'd)

(V) Historical Control. The results of treatment with the test drug are compared with experience historically derived from the adequately documented natural history of the disease or condition, or from the results of active treatment, in comparable patients or populations. Because historical control populations usually cannot be as well assessed with respect to pertinent variables as can concurrent control populations, historical control designs are usually reserved for special circumstances. Examples include studies of diseases with high and predictable mortality (for example, certain malignancies) and studies in which the effect of the drug is self-evident (general anesthetics, drug metabolism).

Historical Control (External)

Retrospective

Unblinded

Selection bias very hard to avoid

Past experience, other non-random experience

Baseline (patient as own) control is a kind of historical control (assume what would have happened).

Historical Controls

Critical Reference -

Sacks, Chalmers, Smith

Am J. Medicine (1982); 72:233-240.

Comparison of RCTs and HCTs for same disease

<u>Always</u>

1. RCT less favorable than HCT
2. Reason was that the historical control was worse than the randomized control (selection bias)
3. Not possible to “adjust” the difference

Many examples of misleading HCTs; great care in relying on one. Addressed in ICH E-10

Table I - Conclusions of RCTs and HCTs on Six Therapeutic Questions

Question Studied	RCT		HCT			
			All Trials		Matched or Adjusted for Prognostic Factors	
	Effec-tive	Ineffec-tive	Effec-tive	Ineffec-tive	Effec-tive	Ineffec-tive
Cirrhosis with Varices	6	14	12	6	2	1
Coronary Artery Surgery	1	7	16	5	9	1
Anticoagulants for Acute Myocardial Infarction	1	9	5	1	3	1
5-FU Adjuvant for Colon Cancer	0	5	2	0	2	0
BCG Adjuvant for Melanoma	2	2	4	0	4	0
DES for Habitual Abortion	0	3	5	0	1	0
TOTALS	10	40	44	12	21	3

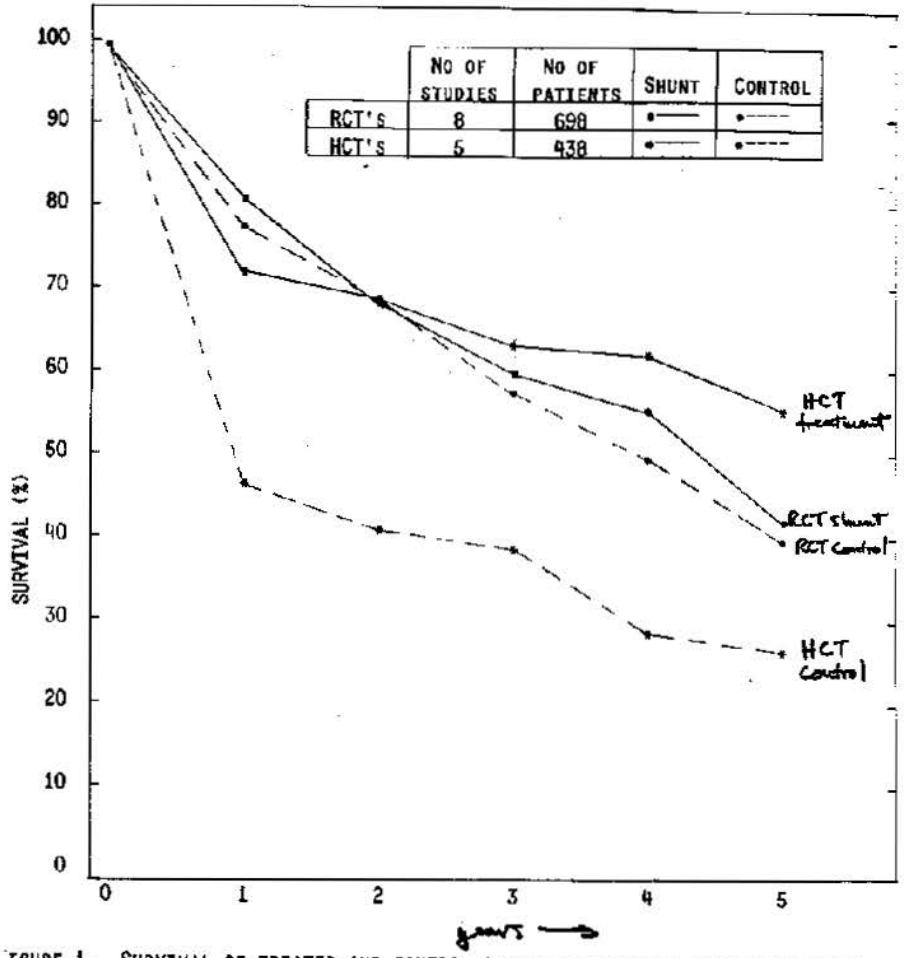


FIGURE 1. SURVIVAL OF TREATED AND CONTROL GROUPS IN CLINICAL TRIALS OF SHUNT SURGERY FOR CIRRHOSIS WITH ESOPHAGEAL VARICES.

Table IV Pooled Survival in Clinical Trials of Medical Versus Surgical Treatment of Coronary Artery Disease

	No. Studies	No. Patients	Percent Survival			
			1 YR	2 YR	3 YR	4 YR
RCT	9	18,861				
Surgical			92.4	89.6	87.6	85.3
Medical			93.4	89.2	83.2	79.8
HCT	6	9,290				
Surgical			93.0	92.2	90.9	88.3
Medical			83.8	78.2	71.1	65.5
Surgical Adjusted*			93.7	92.5	91.2	87.4
Medical Adjusted*			88.2	82.2	70.9	67.7

*Adjusted to have the same proportion of patients with one-, two- and three-vessel disease as in the RCTs.

Historical Controls

Fulminant Hepatitis B - Australia AG Treatment

History

Gocke observed 9 consecutive cases of acute fulminant hepatitis, all fatal despite exchange Tx, steroids, supportive care

Then, 8 hepatitic coma patients given same Rx plus anti-Australia antigen serum, with 5/8 survival

Considered accepting data but concluded it could represent better care, earlier Rx

Therefore RCT in severe hepatitis. Hyperimmune globulin vs. normal serum globulin

[Result - no effect]

Adequate and Well-Controlled Studies (Cont'd)

(3) The method of selection of subjects provides adequate assurance that they have the disease or condition being studied, or evidence of susceptibility and exposure to the condition against which prophylaxis is directed.

Adequate and Well-Controlled Studies (Cont'd)

(4) The method of assigning patients to treatment and control groups minimizes bias and is intended to assure comparability of the groups with respect to pertinent variables such as age, sex, severity of disease, duration of disease, and use of drugs or therapy other than the test drug. The protocol for the study and the report of its results should describe how subjects were assigned to groups. Ordinarily, in a concurrently controlled study, assignment is by randomization, with or without stratification.

Bias reduction <u>before</u> the trial.

Adequate and Well-Controlled Studies (Cont'd)

(5) Adequate measures are taken to minimize bias on the part of the subjects, observers, and analysts of the data. The protocol and report of the study should describe the procedures used to accomplish this, such as blinding.

Bias reduction during after the trial

Minimization of Bias

What can make a well-designed study give the wrong answer:

1. Non-comparability of groups
   - random differences at baseline (bad luck)
   - post-randomization differences
     unavoidable (drop-outs)
     avoidable (bias, unblinding)

2. Analytic bias or failure to correct the analysis appropriately for multiplicity, including:
   1. Exclusions of patients who were randomized - planned vs. unplanned; effect known or not known
   2. Multiple comparisons: multiple endpoints, multiple subsets, grouping of endpoints: planned vs. unplanned
   3. Post-hoc changes in analysis based on knowledge of the results

Minimization of Bias

Comparability of groups

Both before and after start of study

1. Before: well understood; use randomization
   - Demography
   - Disease severity, risk factors
   - Other treatment
   - Study site
   - Concomitant illness

Comparability

2. During study: not as well appreciated, use blinding

- Frequency of visits
- Added treatments
- Patient hopes - placebo response
- Investigator attitude
  - Search for ADRs; attribution of ADRs
  - Compliance; keeping in study
  - Interpretation of an outcome (AMI, yes or no; cause of death, reason for leaving study) - ART
  - Encouragement to perform
  - Exclusion of patients - ART
  - Eligibility
- Differential drop-outs
- Referral of events for blinded adjudication

Unbiased Analysis

1. Multiplicity

Basic problem: Test 2 independent endpoints at p=0.05 (heart attack, stroke), or two subsets at p=0.05 (men, women), the chance of failing to show a difference by chance alone is 0.95 for each one.

Chance of failing to show either is 0.95 x 0.95=0.9, or of showing at least one is 0.1. The chance of showing at least one "significant" finding by chance alone is thus not 0.05 or 1 in 20, but 0.1 or 1 in 10.

Multiple comparisons need statistical correction.

Similar problems with multiple statistical analyses and multiple looks at data.

2. Unbiased Analysis

You can't look at the results and develop a new, not previously planned, analysis.

Lee, et.al.

Subgroup with 3-vessel disease and abnormal contracting ventricle (N=397)

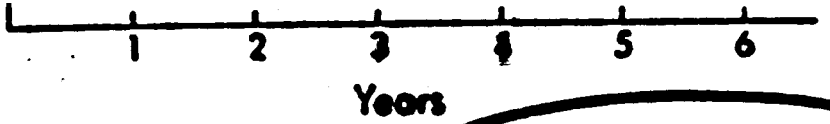


FIGURE 1. *Comparison of long-term survival in a subgroup composed of 194 group 1 patients ("treatment" A) and 203 group 2 patients ("treatment" B) with three-vessel*

A vs. B

$p<0.025$

Unbiased Analysis

1. State analysis plan before study - identify all deviations, changes made prior to unblinding

   GREAT CARE with UNPLANNED ANALYSES

2. Do at least one analysis using all patients (no exclusions).

3. Identify primary endpoints before study and correct/adjust for multiple endpoints.

4. Plan for multiple (interim) looks at data if desired and make statistical correction.

Anturane Reinfarction Trial

Late 1970's RCT Sulfinpyrazone vs placebo in patients 25-35 days post AMI.

Reported near-significant mortality effect and significant effect on early (6 months) and especially sudden cardiac death.

But it was all wrong because

1. Cause-specific mortality was unreliable. Sudden death and AMI and "other" often had the same description; choices called death, "sudden" when in placebo group and the same deaths "MI" or "other" on Anturane.
2. Six deaths in patients randomized to Anturane were dropped after the fact, when they were found "ineligible".

TOTAL CARDIAC DEATHS

	P1	S
A.R.T.	62	43
POOR COMPLIANCE	1	2
LATE INELIGIBLE	0	6
LESS THAN 7 DAYS	5	4
INELIGIBLE <7D	1	0
TOTAL	69	55
		p=~0.2
LATE DEATHS	13	10
TOTAL	82	65
		p=0.162

ART - Conclusions/Lessons

1. Cause of death analyses (cause-specific mortality) is treacherous. We now:
   - have a strong bias toward all-cause mortality
   - sometimes accept CV mortality (but without trying to distinguish further)

2. Pay very close attention to the planned analysis, with great reluctance to look at time or outcome subsets not planned and not accounted for in statistical plan.

3. Insist on full accounting of all randomized patients and an ITT analysis (even if sponsor prefers another).
4. This is all written up [Temple and Pledger, N Engl J Med. 1980 Dec 18;303(25):1488–1492.]

Endpoints of Trials

The choice of study endpoints is critical to drug assessment, but law and regulations say little about it. The endpoint must be clinically meaningful (Court) but can be

- important outcome: death, AMI
- symptom
- surrogate endpoint:

  A surrogate endpoint, or “marker,” is a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint that is a direct measure of how a patient feels, functions, or survives and that is expected to predict the effect of the therapy

Accelerated Approval (21 CFR 314.500)

Nothing in law forbids use of a surrogate endpoint for approval and some are considered valid and regularly used (BP, BS, cholesterol)

But experience with antiarrhythmics, inotropic drugs for heart failure, and more recently experience with toceptrapib (raises HDL cholesterol) has led to considerable skepticism

A rule (1992) on “Accelerated Approval” addressed this, reflecting both skepticism and the sense of urgency that can arise in relation to serious, untreatable illnesses [Incorporated into FDAMA, 1997]

Accelerated Approval

Approval based on a surrogate endpoint “that is reasonably likely, based on epidemiologic, therapeutic, pathophysiologic, or other evidence to predict clinical benefit”.

Conditions:

1. Serious or life-threatening illness
2. Meaningful therapeutic benefit over existing treatments
3. Requirement to study the drug post-approval to “verify and describe its clinical benefit”.
4. Easy removal

Used principally for AIDS drugs (viral load, T4 lymphocytes) and oncologic drugs (response rate in refractory disease)

How Many Studies?
or
When Can an Effectiveness Conclusion be Based on a Single Study

Guidance: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products (May 1998)

Response to FDAMA (1997), (though had been under development for several years), which explicitly allowed approval based on a single study with “confirmatory evidence”

Non-Inferiority Studies

Active control studies, including non-inferiority studies, are an accepted basis for approval (a showing of effectiveness) but as noted earlier, the regulations identify a particular concern: knowing that the active control was effective, and what the effect size was, in the new study (without a placebo group to tell you).

Non-Inferiority Studies - Why?

The principal reason for using an active control non-inferiority design is the inability to use a placebo control because it would be unethical to deprive patients of established important therapy.

Apart from the ethical reason, growing interest in comparative data has led to great interest in active control comparative trials, but if comparative effectiveness is of interest, and a placebo is ethical, you should use a 3 arm (test, control, placebo) study.

Evidence of Effectiveness

There are two distinct approaches to showing effectiveness:

1. Difference-showing

   Superiority of test drug to some control (placebo, active, lower dose) demonstrates drug effect (and assay sensitivity, the ability of the trial to detect differences when they are present). Lack of assay sensitivity, does not lead to an erroneous conclusion that the drug is effective, although it could lead to missing an effective drug

2. Equivalence or non-inferiority in an active control study

   Non-inferiority trials show that the new drug is not worse than the control by a defined amount, the non-inferiority margin M. M must be no larger than the whole effect of the control, i.e., the effect the active control would be expected (<u>known</u>, really) to have in the study. This is the largest possible non-inferiority margin, M_1 and it shows the test drug has some effect. Usually, the margin is smaller than that, M_2, and is no larger than a clinically meaningful difference.

The Logic Is Not The Problem

Showing equivalence to a known active drug that was in fact active in the study would be a sensible way to demonstrate effectiveness

But you can't really show equivalence (except by being superior), so we seek Non-Inferiority,

a misnomer

Really it is showing inferiority of no more than a specified margin M

$$C-T<M$$

So it's really a "not-too-much-inferiority" trial

[Old, naïve way (but still seen in current publications) was to compare C and T, find "no significant difference" and declare victory. A major problem with this, apart from assay sensitivity, was that increasing variance alone (e.g., by having too small a study) will create "success" (no significant difference]

Clinical Trials: Difference-Showing vs Equivalence

Most controlled trials are placebo or no-treatment controlled, and have as a null hypothesis that the effect of the test drug (T) is ≤ 0 (placebo).

$$Ho: T \leq P, \text{ or } T\text{-}P \leq 0$$
$$Ha: T > P, \text{ or } T\text{-}P > 0$$

The alternative is established by showing that the 97½ one-sided lower bound of the CI for T-placebo is > 0.

A successful difference showing trial demonstrates an effect, so long as the defeated control is not < zero. (Easy for a placebo).

Clinical Trials: Difference-Showing vs Equivalence

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, i.e., is more inferior than we are willing to accept.

Ho: C-T ≥ M (T is more inferior to C than M)

Ha: C-T < M (T is less inferior to C than M)

For the study to show that T has <u>any</u> effect, M can be no larger than the whole effect of C in that study, frequently referred to as M_1. Again you compare the 97½% CI upper bound of C-T with M. If you reject the null hypothesis, then T has <u>some</u> effect (> 0).

The smaller M is, the harder it is to show that C-T is < M, because the allowable difference between control and test shrinks. This creates an incentive to choose a higher value for the M.

M is Crucial

Everything depends on the validity of M; if M is larger than the actual effect of C in the study, e.g., if C had no effect in that study, you will reach an erroneous conclusion that T is effective.

M thus needs to be chosen conservatively. If, e.g., you say M=10, then if C-T (97½% CI upper bound) is < 10, say 8, T has an effect. But if in the study the effect of C is in fact 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 regulators to conservative choices of M, with the consequence of large sample sizes.

Study Outcomes

The NI study is intended to show that there is <u>some</u> effect of T. If the control has an effect of M in the study, then consider 3 possibilities:

1. T > C (new drug is better than C). Then M is irrelevant; it's a superiority finding

2. C-T > M_1 (the test drug is more inferior than M_1, the whole effect of C)

The study does not show that T has any effect

3. C-T < M_1

If C had an effect at least as large as M_1 in this study, and if there was assay sensitivity (i.e., if the control really did have an effect of at least M_1), then T has some effect.

What's the Problem

If the logic of the NI study is OK, what's the problem?

The problem is that unlike a finding of superiority, which "speaks for itself," a finding of non-inferiority depends absolutely on an assumption rather than on a measurement.

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

. 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. 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

Fundamental Logic of Trials

Superiority = Efficacy

(if control $\geq$ placebo)

Non-inferiority $\neq$ Efficacy

(unless assay sensitivity is present)

The Assay Sensitivity Problem

I remember exactly when I realized there was a problem, my epiphany: we saw proposed trials in 1978 or so that were going to compare nadolol with propranolol in angina, without any placebo. But we knew the large majority of placebo-controlled propranolol trials had failed (not shown any effect)

So, how could a finding of no difference between N & P mean anything at all?

It couldn't

Problems of Active Controlled Trials

As early as 1982, proposed 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 25 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

Assuring Assay Sensitivity In Non-Inferiority Trials - the Major Problem

In a non-inferiority trial, assay sensitivity is not measured in the trial. That is, the trial itself does not show the study's ability to distinguish active from inactive therapy. Assay sensitivity must, therefore, be deduced or assumed, based on 1) historical experience showing sensitivity to drug effects, 2) a close evaluation of study quality and, particularly important, 3) the similarity of the current trial to trials that were able to distinguish the active control drug from placebo.

In many symptomatic conditions, such as depression, pain, allergic rhinitis, IBS, angina, the assumption of assay sensitivity cannot be made. Trials of effective anti-depressants, e.g., fail to distinguish drug from placebo about half the time.

Assay sensitivity can be measured in an active control trial if there is an "internal standard," a control vs placebo comparison as well as the control vs test drug comparison (i.e., a three-arm study).

Lou Lasagna, 1979

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

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 outcome studies, results are also inconsistent, notably survival post-MI with beta blockers or aspirin. Recent assessments have shown that placebo-controlled trials do not reliably show effects of antibiotics in otitis media, sinusitis, or acute exacerbations of chronic bronchitis.

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

Determining Assay Sensitivity

1. HESDE

In many cases of symptomatic conditions, it is not possible to conclude there is historical evidence of sensitivity to drug effects (and thus potential assay sensitivity for any given trial) because failure to distinguish drug from placebo in what seem to be well-designed studies is not uncommon. A finding of “non-inferiority would therefore be meaningless

YOU CANNOT USE A NI STUDY IN THOSE CASES

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 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

Active Controls Equivalence Credible

1. Some bacterial infections
2. Thrombolytics
3. Treatment of deep vein thrombosis
4. Many stages of HIV infection
5. Treatment of highly responsive tumors (ALL, testicular tumors, ovarian)
6. Anesthetic agents
7. Beta-agonists in bronchospasm
8. Comparison of anticoagulants in chronic AF

Four Critical Steps in Using a Non-Inferiority Design

1. Determining that historical evidence of sensitivity to drug effects exists

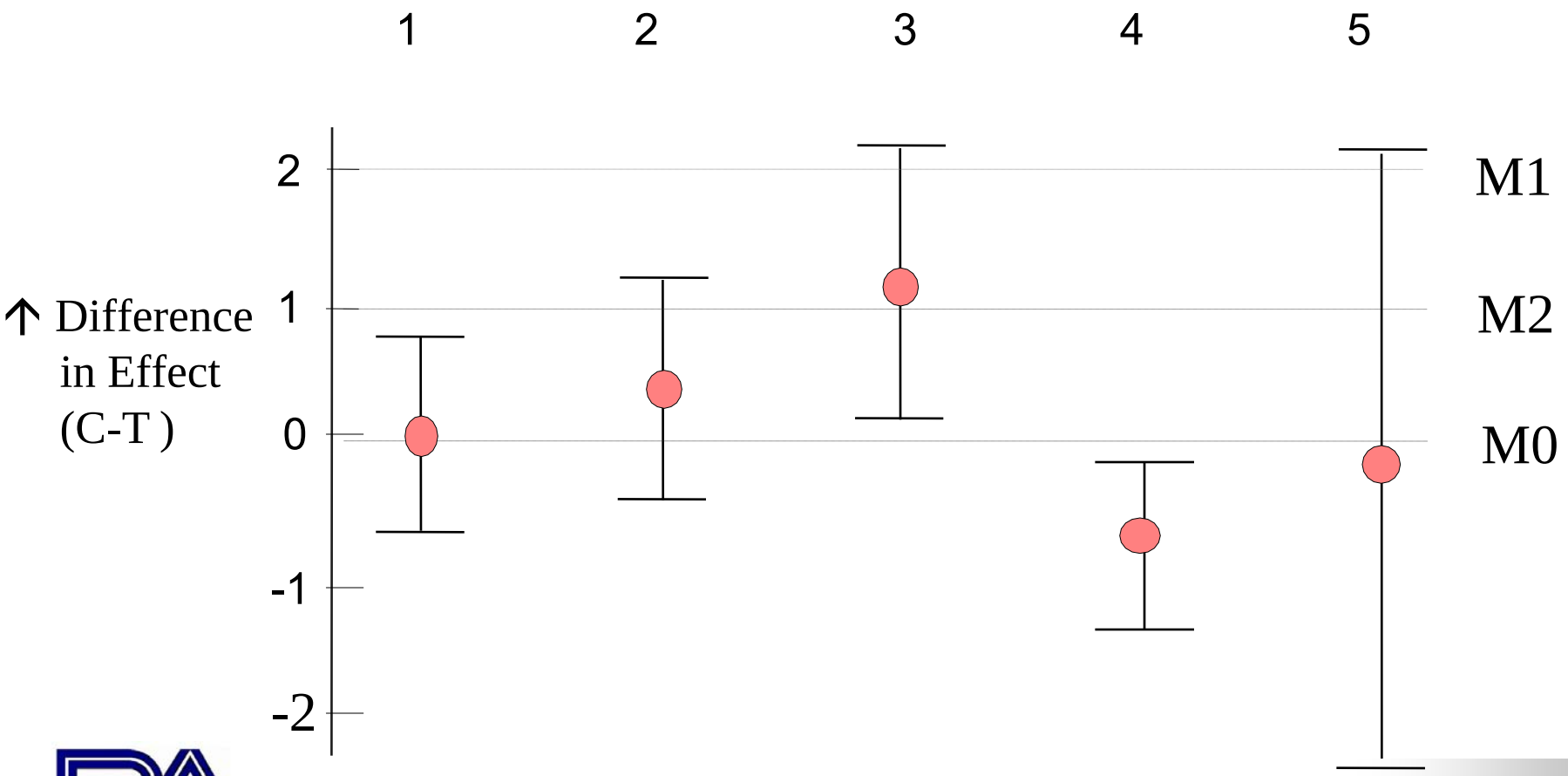
2. Setting an acceptable non-inferiority margin, M_1, a margin no larger than the effect the control can be reliably presumed to have had in the study, and that also reflects the fraction of the control effect that is considered clinically essential, M_2

3. Designing a trial (study population, concomitant therapy, endpoints, run-in periods) that is very similar to the trials for which historical sensitivity to drug effects has been determined

4. Conducting the trial properly and similarly to the historical controls

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 <u>any</u> of the control effect may not be adequate. It is therefore common to choose M_2 as the non-inferiority margin, where M_2 is smaller than M_1 and represents the largest part of the effect of the control (M_1) that can be lost (often chosen as a fraction of M_1). Note that you cannot assure true equivalence or no inferiority at all except by having T be superior to C


1
2
3
4
5
2
1
0
-1
-2
↑ Difference in Effect (C-T)
M1
M2
M0

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 (the clinically acceptable difference) 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 the M_2 chosen is larger than M_1. 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 of these studies did not have a known effect as large as 20% better 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 had been documented.

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

Find M1, First

To design a non-inferiority study, find M_1, the effect you are <u>sure</u> the control agent had <u>in the study</u>. Then if

$$C-T<M_1,$$

the test drug would have at least <u>some</u> effect

<u>Then</u> decide how much of M_1 you need to preserve, (or are willing to lose). If, e.g., you believe that at least 75% of the control effect must be assured, then M_2 can be no greater than 25% of the estimated effect of the control. In that case, M_2 will be the reference for the null (really non-inferiority) hypothesis and the study will need to show

$$C-T<M_2$$

Enrichment

We don't do clinical trials in a random sample of the population. We try to make sure people have the disease we're studying (entry criteria), have stable disease with stable measurements (lead in periods), do not respond too well to placebo (placebo lead in periods), have disease of some defined severity, and do not have conditions that would obscure benefit. These efforts are all kinds of ENRICHMENT, and almost every clinical trial uses them. There are, in addition, other steps, not as regularly used, that can be taken to increase the likelihood that a drug effect can be detected (if, of course, there is one).

Enrichment

Enrichment is prospective use of any patient characteristic – demographic, pathophysiologic, historical, genetic, and others – to select patients for study to obtain a study population in which detection of a drug effect is more likely.

This occurs to a degree in virtually every trial, although enrichment may not be explicit, and is intended to increase study power by:

- Decreasing heterogeneity
- Finding a population with many outcome events, i.e., high risk patients – prognostic enrichment
- Identifying a population capable of responding to the treatment – predictive enrichment

Enrichment

The increased study power facilitates “proof of principle” (there is a clinical effect in <u>some</u> population) but it can leave open 1) the question of generalizability of the result and how the drug will work in other populations, as well as 2) the question of how much data are needed before or after approval in the “non-selected” group.

Kinds of Enrichment

1. Practical – virtually universal – decrease heterogeneity and "noise"
   - Define entry criteria carefully
   - Find (prospectively) likely compliers (VA HT studies)
   - Choose people who will not drop out
   - Eliminate placebo-responders in a lead-in period
   - Eliminate people who give inconsistent treadmill results in heart failure or angina trials, or whose BP is unstable
   - Eliminate people with diseases likely to lead to early death
   - Eliminate people on drugs with the same effect as test drug

In general, these enrichments do not raise questions of generalizability, although eliminating people who do not tolerate the drug might do so.

Kinds of Enrichment (cont)

Apart from practical enrichment strategies fall into two distinct types:

2. Prognostic enrichment - choosing high risk patients, i.e., those likely to have the event (study endpoint) of interest, or likely to have a large change in the endpoint being measured, e.g., a high rate of deterioration.

   This has study size implications, of course, but also therapeutic implications. A 50% change in event rate means more in high risk patients (10% to 5%) than in low risk patients (1% to 0.5%) and could lead to a different view of toxicity.

3. Predictive enrichment - choosing people more likely to respond to treatment.

   Choices could be based on pathophysiology, proteomic/genomic observations, patient history, early response of a surrogate endpoint (e.g., tumor response on some radiographic measure), or a history of response.

Past Selection of High Risk Patients (more likely to have events)

Although the information distinguishing individuals is growing exponentially, we've had such information before

- Epidemiologic risk factors for CV events
  - Cholesterol, blood pressure levels
  - Diabetes
  - Prior events (AMI, stroke, PVD)
  - Family history
  - Gender, race, age
- Individual measurement/history in various settings
  - Previous breast cancer to predict breast Ca
  - Tumor histology to predict metastasis
  - Arteriogram, echocardiogram, exercise testing to predict CV events
  - Evidence of MCI as predictor of Alzheimer's Disease; genetic predictors

Past Selection of Responders-Pathophysiology

Pathophysiological selection, based on understanding of disease, is common and usually represents an attempt to select potential responders.

- Edema can result from hepatic, renal or cardiac causes. Choose the last for study of an inotrope or other cardiac intervention

- CHF can result from systolic or diastolic dysfunction. Choose the former for study of a positive inotrope, the latter for a CCB. With other kinds of drugs, e.g., diuretics or ACEIs, might stratify to see if results differ by pathophysiology

- We distinguish (some) causes of pain: angina, vasospastic angina, migraine, menstrual pain, etc., where we believe etiologies are distinct and particular pharmacologic effects are pertinent. We study each separately but would have interest in how drug works in others

Enrichment – High Risk Patients

1. Oncology

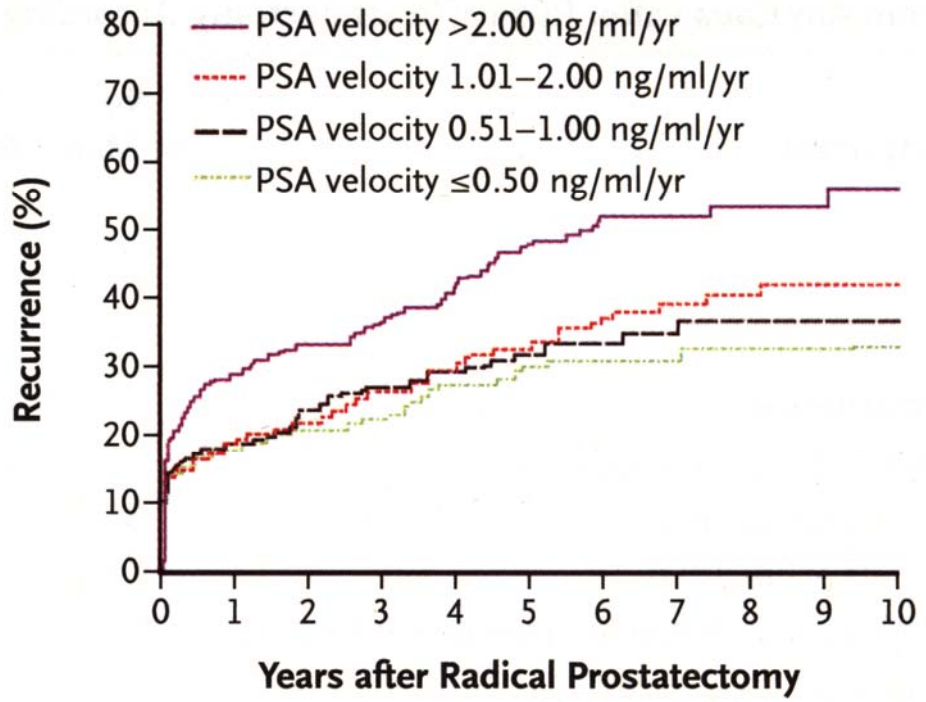
   Tamoxifen prevented contralateral breast tumors in adjuvant setting (very high risk); it was then studied in people with more general high risk. This was needed a) to have enough endpoints to detect a possible effect and b) because of concern about toxicity. It was labeled for the group studied, with access to Gail Model calculator to assess risk. There was no reason in this case to expect larger % effect in the people selected, but more events would be prevented.

Enrichment – High Risk Patients

1. Oncology (cont.)

   Potential selection method for frequent endpoints (not tried yet, to my knowledge)

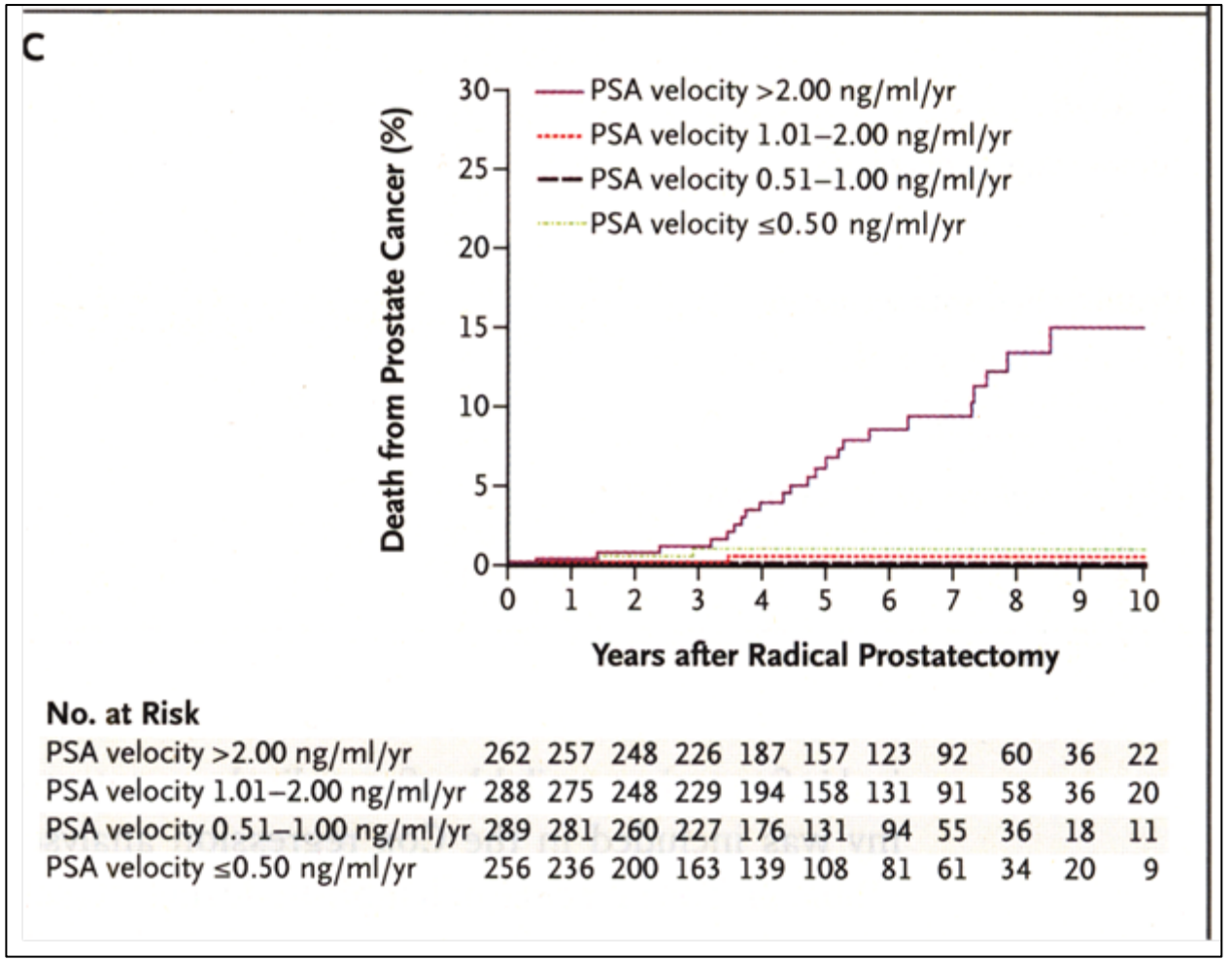
   D'Amico showed [NEJM 2004; 351:125-135] that in men with localized prostate Ca, following radical prostatectomy, PSA "velocity" (PSA increase > 2 ng/ml during prior year) predicted prostate Ca mortality almost 100% over a 10 year period. There were essentially no deaths from prostate Ca (many from other causes), even though recurrence rates were not so different



No. at Risk

PSA velocity >2.00 ng/ml/yr	247	173	155	132	104	81	60	45	31	19	13
PSA velocity 1.01–2.00 ng/ml/yr	280	218	191	167	133	101	84	56	36	19	15
PSA velocity 0.51–1.00 ng/ml/yr	287	226	193	158	120	92	64	36	23	14	9
PSA velocity ≤0.50 ng/ml/yr	249	190	156	128	103	84	58	43	24	13	5

Kaplan-Meier Estimates of Disease Recurrence (Panel A) after Radical Prostatectomy, According to the Quartile of PSA Velocity during the Year before Diagnosis



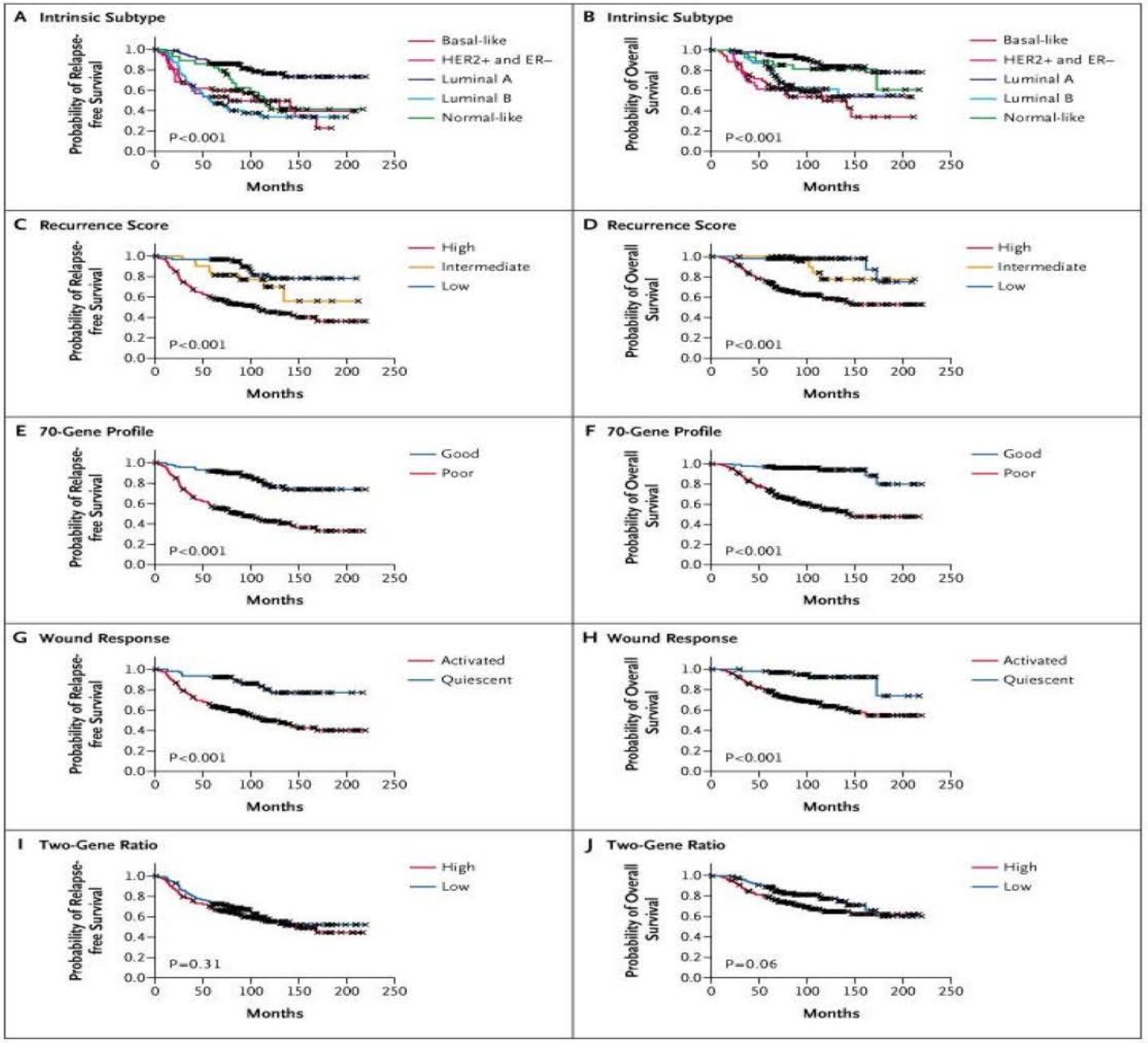
Kaplan-Meier Estimates of the Cumulative Incidence of Death from Prostate Cancer (Panel C) after Radical Prostatectomy, According to the Quartile of PSA Velocity during the Year before Diagnosis

Enrichment – High Risk Patients

1. Oncology (cont)

Fan, et al [NEJM 2006; 355: 560-69] recently applied 5 different gene-expression profiling approaches, intended to predict breast cancer recurrence rates, to a 285 patient sample treated with local therapy, tamoxifen, tamoxifen plus chemo, or chemo alone.

The results and methods used are shown on the next slide. Four of the 5 methods had high concordance and a striking ability to predict outcome and the differences were very large. The implications for patient selection are obvious, whether the endpoint is recurrence or survival. Studies should select poorer prognosis patients to have a better chance of showing a drug effect.


A Intrinsic Subtype
Probability of Relapse-free Survival
Basal-like
HER2+ and ER–
Luminal A
Luminal B
Normal-like
P<0.001
Months
B Intrinsic Subtype
Probability of Overall Survival
P<0.001
C Recurrence Score
High
Intermediate
Low
D Recurrence Score
E 70-Gene Profile
Good
Poor
F 70-Gene Profile
G Wound Response
Activated
Quiescent
H Wound Response
I Two-Gene Ratio
P=0.31
J Two-Gene Ratio
P=0.06

Enrichment-High Risk Patients

2. Cardiovascular

Long routine to choose patients at high risk (secondary prevention, post-AMI, or stroke, very high cholesterol, very Severe CHF, undergoing angioplasty) so there will be events to prevent. For example

- CONSENSUS (enalapril) in NYHA class III-IV patients studied only 253 patients, showing dramatic survival effect in only 6 months study. Mortality untreated was 40% in just 2 months, and treatment showed a 40% reduction. Later studies needed many 1000's of patients
- First lipid outcome trial (4S - Simvastatin) in a post-MI, very high cholesterol population: 9% 5 year CV mortality
- JUPITER study of rosuvastatin included people with "normal" LDL but high CRP

Selection of High Risk Patients

3. Other

Identifying people at high risk is especially important in “prevention” or risk reduction efforts. Apart from the CV risks we know about, there may be genetic predictors of risk (e.g., for Alzheimer’s Disease or particular cancers) or early clinical signs (people with minimal brain dysfunction in prophylaxis of Alzheimer’s Disease) or biomarkers (amyloid in brain as predictor of Alzheimer’s Disease.

Selection of Likely Responders

Identifying the people who will respond to a treatment, then formally studying them, greatly enhances the power of a study, facilitating approval, and also may have implications for how a drug will be used.

It can be especially critical when responders are only a small fraction of all the people with a condition, e.g., because they have the “right” receptor. In such a case finding a survival effect in an unselected population may be practically impossible.

Sometimes selection is based on understanding of the disease, i.e. pathophysiologic selection, and seems obvious, as the examples earlier showed.

Selection of Likely Responders

- Hypertension can be high-renin or low-renin. High renin population would show a much larger effect than a mixed population to ACEIs, AIIBs, or BBs.

- We study antibiotics in bacterial infections sensitive to the antibacterial

- A well-established genetically determined difference could be the basis for a pathophysiologically selected population. A marker associated with a particular tumor characteristic could be a basis for selection. Most convincing so far are tumor genetics: Herceptin for Her2+ breast tumors; selection of ER^+ breast tumors for anti-estrogen treatment.

Selection of Likely Responders

Even if pathophysiology is unclear, likely responders could be identified by an initial short-term response. There is a history of this:

- CAST was carried out in people who had at least a 70% reduction of VPB's. Only "responders" were randomized.
- Trials of topical nitrates were carried out only in people with a BP or angina response to sublingual nitroglycerin.
- Anti-arrhythmics were developed by Oates, Woosley, and Roden by open screening for response, then randomizing the responders.
- Every randomized withdrawal study has this characteristic.

Selection of Likely Responders

Selection could be based on response of a biomarker; that is, study the entire group and randomize only those with a good response. Possibilities

- Tumor that shows early metabolic effect on PET scan
- Tumor that shows early response on blood measure (PSA)
- Tumor that doesn't grow over an n-week period (it would be hard to randomize tumor responders to Rx vs. no Rx)
- Only patients with LDL effect > n (or some other less studied lipid)
- Only patients with CRP response > x

Selection of Likely Responders

We are at the very beginning of searching for genetic or other characteristics that will predict response. These could be pathophysiologic, that is, based on understanding of disease or drug mechanism (role of her 2 receptor in response to Herceptin; role of EGFR in response to erlotinib), generally with these factors identified prospectively, and with patients either selected by, or stratified by, that factor. But the selection could be simply empirical or descriptive: run a trial in unselected patients with depression, bipolar disease, lipid abnormalities, heart failure and link a genetic baseline finding with response. In fact, one could search widely for such a relationship. The usual course would then be to study the genetically described subset prospectively. Tarceva data illustrate the potential. (I should acknowledge some recent uncertainty about some of the measurements and note that this was not prospectively planned).

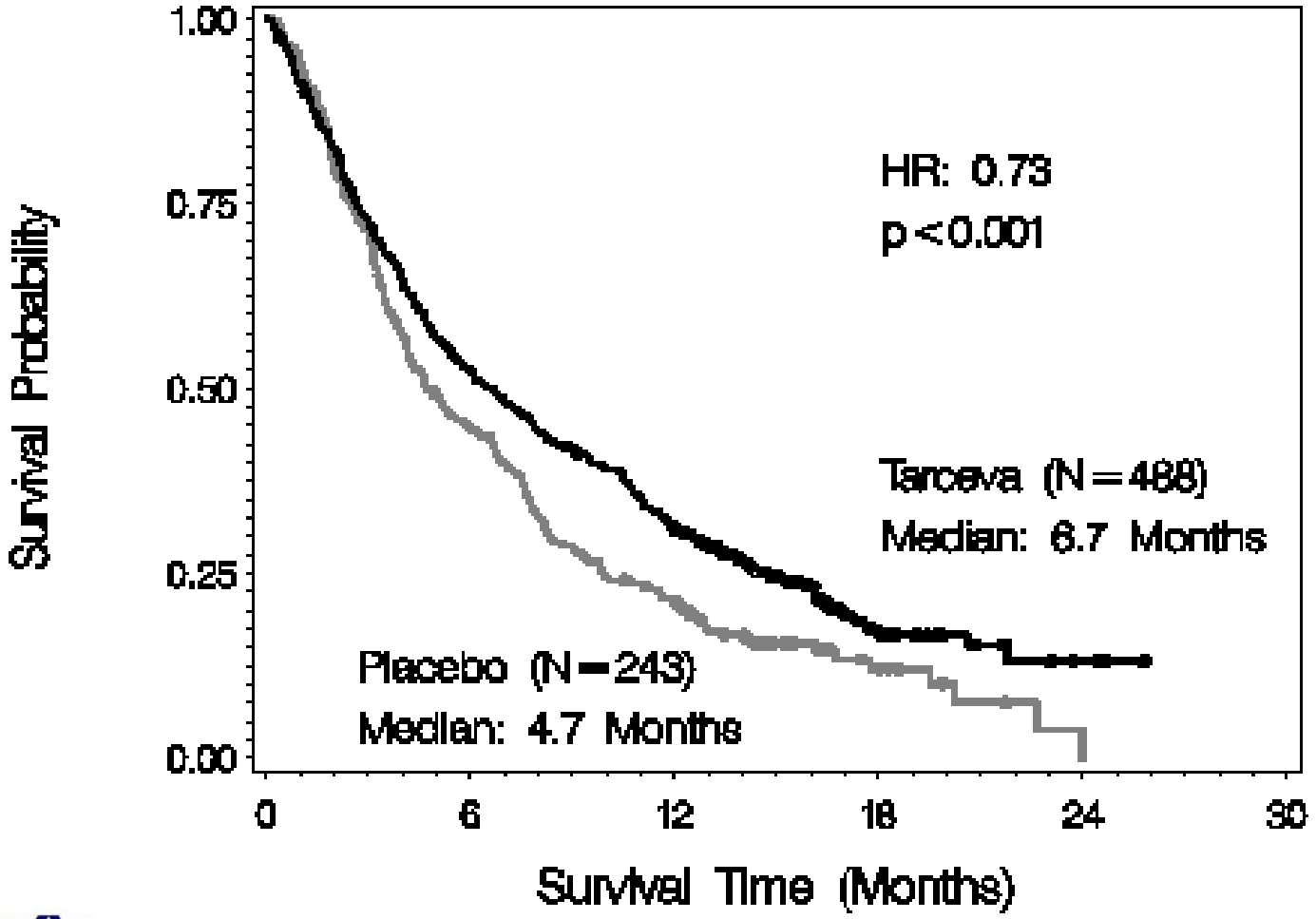
Selection of Likely Responders

Tarceva (erlotinib)

Randomized, DB, placebo-controlled trial of Tarceva 150 mg in 731 patients with locally advanced or metastatic NSCLC after failure of ≥ 1 prior regimen. Randomized 2:1 (488 Tarceva, 243 placebo). Study overall showed clear survival effect

	Tarceva	**Placebo**	**HR**	**CI**
survival (mos.)	6.7	4.7	0.73	0.61-0.86 $p<0.001$
1 year survival	31.2%	21.5%		

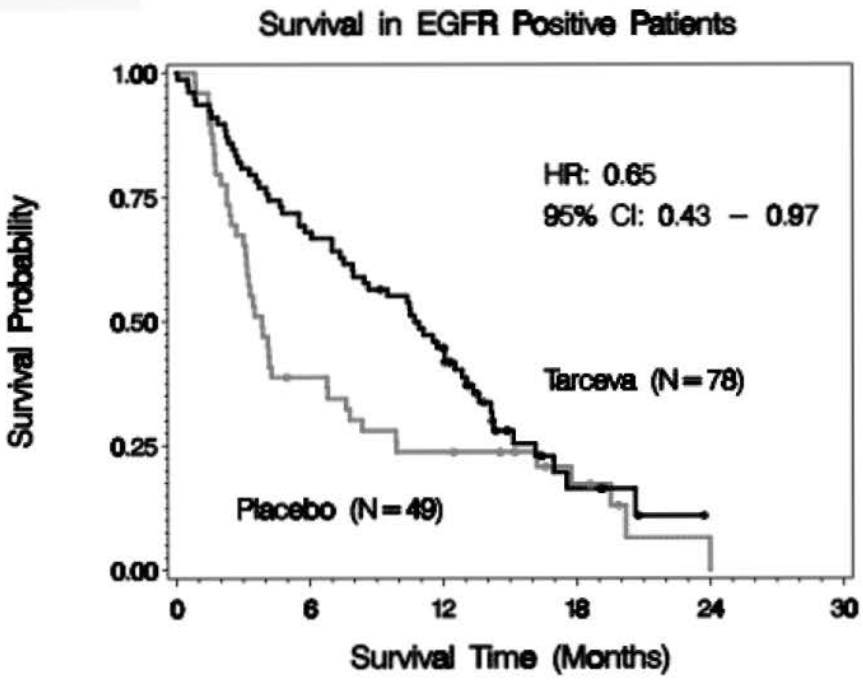
Kaplan-Meier Curve for Overall Survival of Patients by Treatment Group

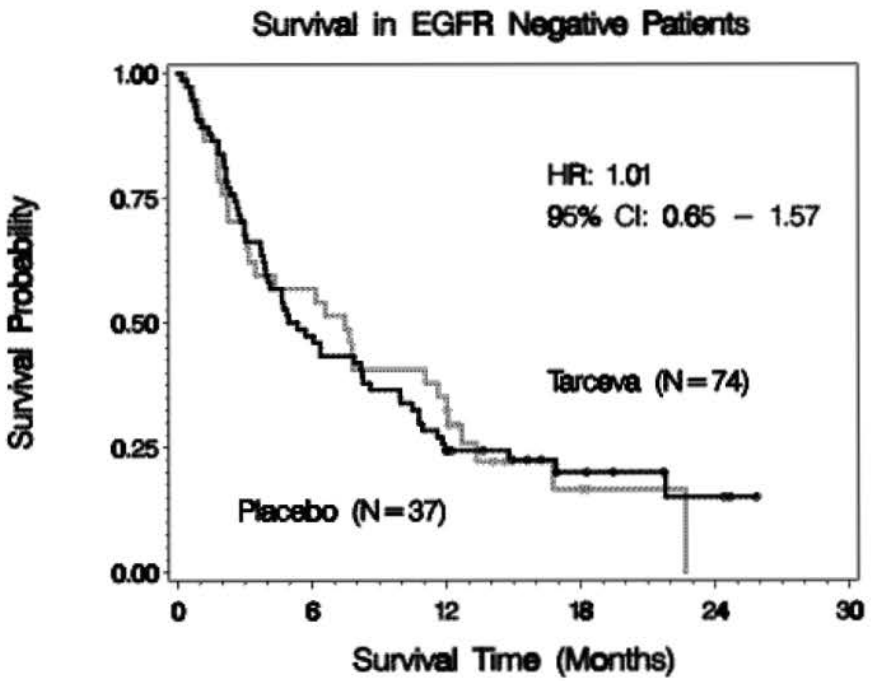


Tarceva (erlotinib)

Tumors were examined for EGFR expression status in 238 (of 731) patients. EGFR+ was defined as ≥10% staining using DAKO EGFR pharmDx kit.

	Tarceva	Placebo	HR	CI
EGFR+ (127) Survival (mos)	78 10.71	49 3.84	0.65	(0.43-0.97) p=0.033
EGFR- (111) Survival	74 5.35	37 7.49	1.01	(0.65-1.57) p=0.958


Survival in EGFR Positive Patients
Survival Probability
1.00
0.75
0.50
0.25
0.00
HR: 0.65
95% CI: 0.43 – 0.97
Tarceva (N=78)
Placebo (N=49)
0
6
12
18
24
30
Survival Time (Months)


Survival in EGFR Negative Patients
Survival Probability
1.00
0.75
0.50
0.25
0.00
HR: 1.01
95% CI: 0.65 – 1.57
Tarceva (N=74)
Placebo (N=37)
0
6
12
18
24
30
Survival Time (Months)

Randomized Withdrawal

Amery in 1975 proposed a “more ethical” design for angina trials, which then often ran 8 weeks to 6 months in patients with frequent attacks (before regular CABG and angioplasty).

Patients initially receive open treatment with the test drug, then are randomized to test drug (at one or more doses) or placebo. Endpoint can be time to failure (early escape) or conventional measure (attacks per week).

These trials are all enriched with people doing well on treatment. Also, no new recruitment is needed.

Patients on treatment with sodium oxybate for cataplexy with narcolepsy for 7-44 months randomized to continued treatment of placebo

	median attacks/2 weeks	
	Baseline	Change in Rate
Placebo (29)	4.0	+21.0
sodium oxybate (26)	1.9	0

$p<0.001$

Clearly demonstrated persisting long-term effect and provided a confirmatory trial in practically no time.

Randomized Withdrawal

An extremely effective design in many cases, as it includes patients who tolerate the drug and who appear to benefit. It is an efficient way to document long-term effect without long-term placebo, and is widely used:

- To show long-term prevention of recurrent depression (studies invariably successful in contrast to 50% failure rate in acute depression).
- To show long-term BD effect in hypertension (long-term placebo would be unethical)

Potential use whenever drop-outs are a problem (e.g., long-term effect on pain).

Dose Response

What could be more obvious in a field allied to pharmacology than finding D/R

Yet until late 1970's, study designs completely ignored this issue and, indeed, almost guaranteed that D/R data would <u>not</u> be developed

Note that a D/R study can be both

1. An adequate and well-controlled study showing effectiveness [21 CFR 314.126]

2. A source of information about D/R

History

Prior to late 1970's

Standard design (except analgesics) was titration to effect or tolerability. Reasons:

- All patients get a dose that should work; more likely to beat placebo; no "waste"
- Seems to match clinical practice
- Seems safe

But, unless analyzed imaginatively (Sheiner, NONMEM), and only in some cases, titration obscures D/R

- Confounds dose and time; especially for safety
- Group on a given dose not randomized to dose
- Typically gives umbrella D/R
- Because many diseases improve with time, creates <u>impression</u> that higher dose is adding to effect

Epiphany and Examples

Chlorthalidone - Materson/Tweeddale

Nadolol - 100-1000 fold excess

Captopril - XS dose gives agranulocytosis

Guanabenz - titration to universal toxicity

Chlorthalidone

Standard dose (same for HCTZ) 100 mg; hypoK recognized but not considered a problem. The 100 mg dose was used in VA studies, HDFP, etc.

Not silly: if you look at Na clearance, it increases to 100 mg and even higher

But no data on D/R for antihypertensive effects or toxicity

Then we saw two studies in CP&T, 1978 using unfamiliar designs, with patients randomized to fixed doses, either parallel (Materson) or x-over (Tweeddale)

TABLE 1

DATA OF MATERSON

Fall in blood pressure (systolic/diastolic) from baseline in erect and supine position with each of four dose levels of chlorthalidone and placebo.

Dose	Fall in Blood Pressure (mmHg)	
	Supine	Standing
Placebo	0/2	0/0
12.5 mg	5/4	6/4
25 mg	11/5	15/7
50 mg	10/6	14/5
75 mg	11/6	14/6

Materson and Tweeddale showed that 25 mg gave the full antihypertensive effect (up to 200 mg by Tweeddale did not add to effect) but higher doses gave

- lower K
- higher BS, cholesterol, UA

A Better Way

Materson and Tweeddale had used a randomized, fixed dose, dose-response design. We noted that this method could teach us about D/R as previous designs could not, and began recommending this design

We were moved to this by a number of examples of major overestimates of the necessary dose

In retrospect, the discovery that only 25 mg of chlorthalidone gave the full antihypertensive effect of a diuretic was critical. Epidemiologic data (Psaty), increased mortality in the 100 mg chlorthalidone group in MRFIT, suggest that the high diuretic dose is directly related to sudden death and that hypokalemia, leading to arrhythmias, probably explains failure of antihypertensive therapy to affect CV mortality much in early studies

Nadolol

Trials in angina and hypertension

Start 40-80 mg/day

Titrate in one week to 240 to 320 mg

It was obvious that even 40 mg had effect, so labeling recommended starting at 40 mg, but still up to 320 mg; that dose is recommended in current labeling

In fact, ED50 is 0.3 mg, based on HR effects, so recommended dose is almost 3 orders of magnitude too high

May not matter with nadolol, but could for a cardio-selective BB

Guanabenz - Contrast with Guanfacine

Central alpha agonist approved 1982 for hypertension; ADRs were frequent and appeared typical of class

Studies started at 8 mg b.i.d. and titrated rapidly to 16 or 32 mg daily; ADR rates were far worse than clonidine.

28% dry mouth

39% drowsiness or sedation

17% dizziness

10% weakness

All rates were far greater than placebo and clonidine, a picture of a virtually unusable drug, probably because of bad D/R assessment. Guanfacine, pharmacologically identical, avoided that picture as will be shown, by better dose-finding

Change

Impressed by Materson, who used a design we never saw (but old Dollery paper on guanethidine used it), as well as by the discouraging examples shown, we began asking for

Randomized, parallel, fixed dose D/R study

- Now world norm - ICH E-4
- In FDA regulations since 1985 as a kind of adequate and well-controlled study
- Frequent use, uniformly in hypertension, depression, many other conditions
- Perhaps has suppressed interesting alternatives. Sheiner has shown usefulness of titration designs, properly analyzed

An example: guanfacine

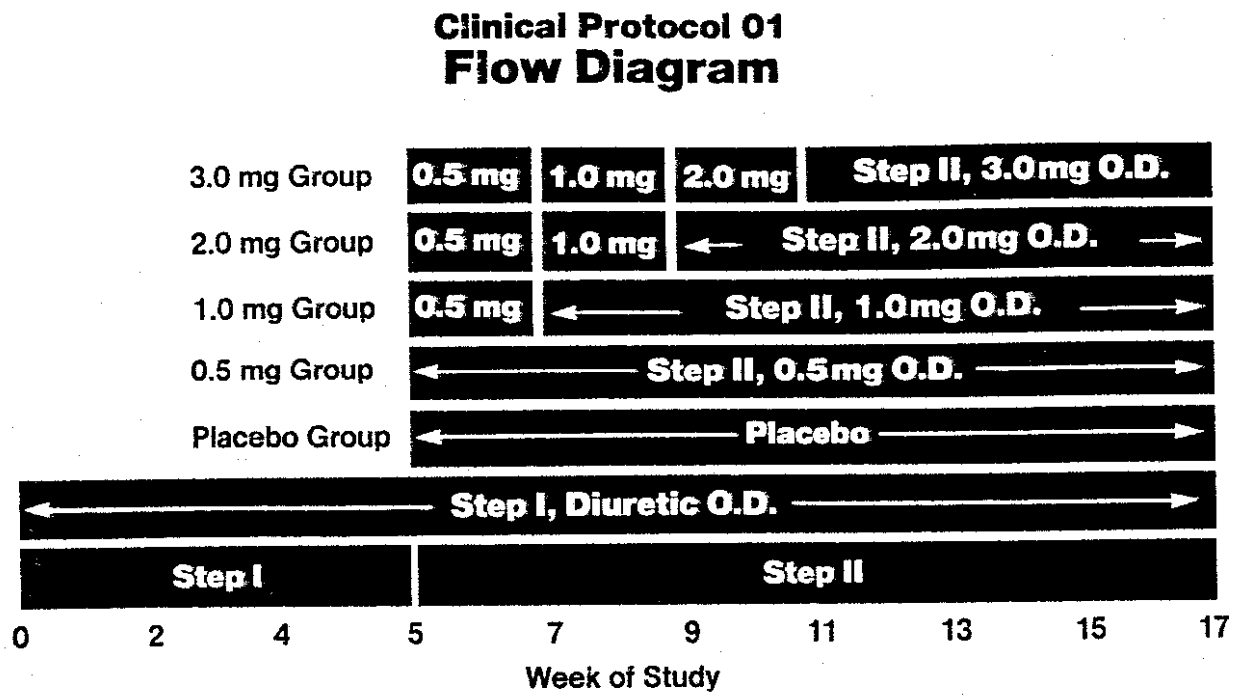


Figure 3 Guanfacine Trial: treatment flow diagram. All patients received diuretic for a 5 week single-blind placebo period (Step I), then were randomized (Step II) into one of the five treatment groups while continuing to receive the diuretic.

Endpoint Means by Treatment Group

Response Criteria	Statistic	Treatment Group				
		Placebo	0.5	1.0	2.0	3.0
	N	63	63	64	58	59
Diastolic Blood Pressure	Mean Change	−7.1	−5.6	−12.7	−13.3	−13.1
Systolic Blood Pressure	Mean Change	−4.7	−4.7	−14.0	−11.6	−15.6
Mean Arterial Pressure	Mean Change	−6.3	−5.3	−13.1	−12.8	−13.9
Heart Rate	Mean Change	+1.2	+2.1	−4.4	−4.7	−4.5

Figure 4 Guanfacine trial: changes from baseline in sitting diastolic, systolic, and mean blood pressure, and heart rate at 12 weeks

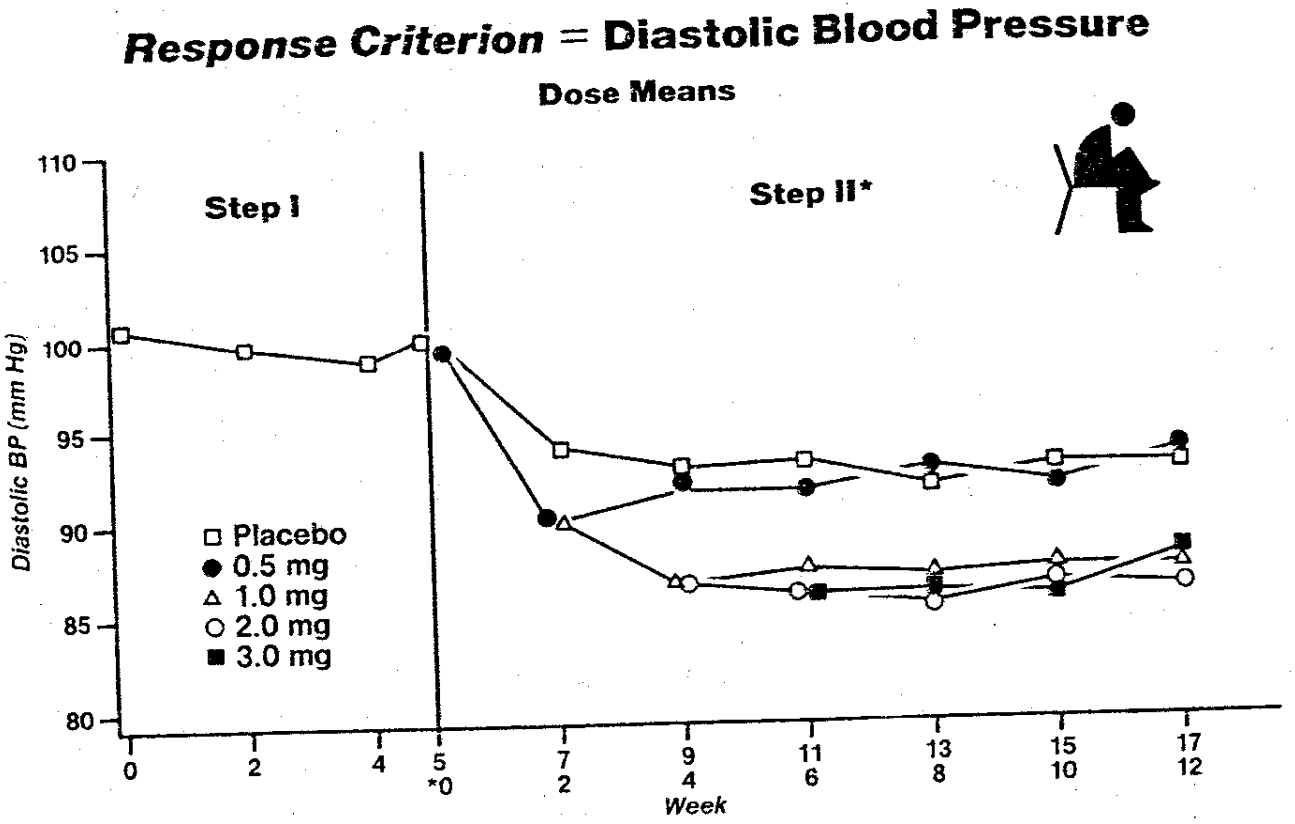


Figure 6 Guanfacine trial: diastolic pressure in all dosage groups in relation to time.

Frequency Distribution of Patients with Most Common Adverse Experiences (Possibly or Probably Related Only)

Adverse Experience	Assigned Treatment Group				
	Placebo	0.5mg	1.0mg	2.0mg	3.0mg
N =	73	72	72	72	72
Dry Mouth	5	4	6	8	20
Somnolence	1	3	0	1	10
Asthenia	1	3	2	6	8
Dizziness	2	1	3	6	3
Headache	3	4	4	1	2
Impotence	1	1	0	2	3

Figure 7 Guanfacine trial: frequency of specific adverse experiences in relation to dose

Risperidone Fixed Dose Studies (8 week BPRS change from baseline)

	Study	
Dose	**024**	**204**
Placebo (n=86)		+2.2
1 mg (n=226)	-6.7	
2 mg (n=87)		-2.9
4 mg (n=227)	-10.2	
6 mg (n=88)		-11.2
8 mg (n=228)	-9.9	
10 mg (n=85)		-5.7
12 mg (n=225)	-9.0	
16 mg (n=85)		-8.5
16 mg (n=223)	-9.7	

Risperidone ADR's

	Dose Group				
ADR	0	2	6	10	16
Parkinsonism Scores	1.2	0.9	1.8	2.4	2.6
EPS Rate	13%	13%	16%	20%	31%

	Dose Group				
ADR	1	4	8	12	16
Parkinsonism Score	0.6	1.7	2.4	2.9	4.1
EPS Rate	7%	12%	18%	18%	21%

STUDY 57-3: BB/Diuretic Factorial Study

- POPULATION: - ESSENTIAL HYPERTENSION
  SIDP 95-115 mm Hg

- DOSES: - BISOPROLOL 0, 2.5, 10 AND 40 mg
  HCTZ 0, 6.25 mg, AND 25 mg
  AND ALL COMBINATIONS

- DESIGN: - MULTICENTER, RANDOMIZED
  DOUBLE-BLIND, 3 X 4 FACTORIAL
  - 4-6 WEEK PLACEBO RUN-IN
  - 12-WEEK DOUBLE-BLIND TREATMENT
  - 2-WEEK TAPER

FDA

STUDY 57-3
OBSERVED MEAN REDUCTION IN SITTING DIASTOLIC BLOOD PRESSURE, mm Hg (±SE)

		BISOPROLOL			
		0 mg	2.5 mg	10 mg	40 mg
	0 mg	3.7 (0.8)	8.1 (0.8)	11.2 (0.9)	12.7 (0.9)
HCTZ	6.25 mg	5.8 (1.4)	10.5 (1.3)	14.3 (1.3)	15.3 (1.2)
	25 mg	9.0 (1.2)	14.0 (1.2)	14.0 (1.2)	16.8 (1.2)

STUDY 57-3
MEAN CHANGE IN SITTING HEART RATE (bpm) BY TREATMENT GROUP

		BISOPROLOL			
		0 mg	2.5 mg	10 mg	40 mg
HCTZ	0 mg	-1.6	-3.8	-10.7	-13.6
	6.25 mg	-0.8	-3.7	-9.8	-14.6
	25 mg	+0.2	-5.7	-9.0	-13.8

STUDY 57-3
ADVERSE EXPERIENCES (%) IDENTIFIED AS SIGNIFICANTLY RELATED TO DOSE

COMBINATION THERAPY - STUDY 57-3

	B/0*	B2.5	B10	B40
n	125	128	129	130
DIARRHEA	0.8	0.8	4.7	6.9
SOMNOLENCE	0	0.8	3.9	3.8
ASTHENIA	0	1.6	0.8	3.8
DYSPEPSIA	0	0	0	3.8

*EACH BISOPROLOL DOSE GROUP REFLECTS TREATMENT WITH PLACEBO, HCTZ 6.25 & HCTZ 25 mg.

FDA

Potassium
Patients Normal At Baseline

	Placebo	Bisoprolol 5 mg	Bisoprolol 5 mg/ HCTZ 6.25 mg	HCTZ 25 mg
Mean Change from Baseline, mEq/L	+0.02	+0.02	-0.08**	-0.31*
Reduction ≥ 0.4 mEq/L, %	5.3	8.8	15.4	33.1
Value < 3.4 mEq/L, %	0	0.7	0.7	6.5

* $p < 0.01$ vs Placebo, B5/H6.25, B5
** $p < 0.05$ vs B5

Lotronex (2001)

More recently, a potential blockbuster drug for IBS associated with diarrhea, Lotronex (alosetron) was removed from the market because of cases of ischemic colitis and severe constipation leading to fatal obstructions. It has been returned under a limited distribution with a reduced dose. The entire phase 3 program used a 2 mg daily dose, producing severe constipation in about 1/3 of patients (this for a treatment for diarrhea). You might have thought they'd have tried dosing 0.5, 1, 2 mg, or dropping back after an initial 2 mg, or going to q 2d treatment, but they didn't.

ICH E4 (1994)

The first FDA general advice on D/R

Strong encouragement to make D/R part of every stage of development and to know shape and location of D/R for favorable and unfavorable effects

Identified randomized, parallel, fixed dose, dose-response study, generally with a placebo group (the studies I used as illustrations; the dose can be titrated to the fixed dose) as the gold standard, but found value in X-over designs, fixed titration designs, optional titration (more later), even encouraging retrospective examination for concentration-response relationships for toxicity

Did not say how to use those findings

ICH E-4

General advice on D/R

1. Strong encouragement to identify whole D/R curve for benefit and toxicity; frequent error is too narrow a range. Placebo desirable.
   - Choose starting dose
   - Identify titration steps
   - Find plateau-dose increase useless
   - Don’t forget dose interval

2. Group values are what we get but individual D/R also of interest. Need to give each person > 1 dose to find this, rarely done

What's the Problem? Where are Improvements?

- More attention in phase 3 to studying the full dose range; phase 2 does not provide adequate information
- Efficiency
- Examine maintenance for dose, dose interval
- Learn about individual D/R, not just group responses

More Attention in Phase 3

Study a full range of doses in phase 3 to establish dose response for both favorable and unfavorable effects and to locate less than fully effective dose that may still be useful. The idea that dose-ranging is a phase 2 activity is totally, and damagingly, wrong. Most phase 3 studies, unless they are very large and it is practically impossible, should include more than one dose and sometimes different regimens (o.d. vs. b.i.d.). The familiar “all-at-once” phase 3 does not easily allow sequential dose-finding but multiple doses can be built in (risperidone). There is no doubt this is a particular problem when a drug shows promise in a bad disease with no good treatment (but recall AZT poor tolerability)

Efficiency

a. Use PD information and efficient designs to narrow range of doses to study clinically:

1) Where PD mechanism is well understood (ACEI's, AIIB's, beta blockers, inhibitors of platelet function) use the PD information, with particular attention to duration, to identify dose range (but don't just believe it; test the expectation; sometimes clinical effect has different duration)

- This is commonly done for ACEI's where effects on ACE activity and AI response are measured, but information is not always used. Captopril, e.g., was used at doses of 150 mg t.i.d., when even a few mg gave substantial (? full) inhibition of ACE); led to early toxicity (agranulocytosis)

Efficiency

b) Consider conducting dose response studies for effectiveness in known responders to the drug or drug class to increase sensitivity, or identify responders pharmacologically, if possible. The only effect of including non-responders is to obscure (flatten) the dose-response relationship. Note, though, that non-responders may have adverse effects and cannot be ignored.

It would usually be important to test non-responders separately to see if they merely have a shift in D/R, an important discovery, if true, and, if responders are not identifiable, studies in a non-selected population would be needed to assess overall B/R.

Parallel Dose-Response Studies of Indapamide

	Study Dose (mg)	n	Baseline BP	Decrease from Baseline (S/D)	
				Standing	Supine
1. Micheal, et al	placebo	17	146/102	3/3	1/1
	1.0	14	143/103	7/5	6/6
	1.5	13	141/101	5/4	5/3
	2.0	15	150/102	21/9*	18/7
	2.5	14	151/104	20/9	17/7
2. Mroczek, et al	placebo	19	153/103	1/2	1/2
	1.0	21	155/104	12/5	10/5
	2.5	21	148/102	14/7	15/6
	5.0	20	153/102	14/6	13/6
3. Sanchez-Torres	placebo	8	163/103	+6/3	+0/6
	1.0	9	174/106	10/4	10/8
	2.5	9	164/104	29/12	22/6
	5.0	8	171/105	37/15	28/15
4. Multicenter vs HCTZ	2.0	30	141/101	12/8	11/7
	2.5	25	147/103	12/7	11/7
	HCTZ 100	28	150/101	12/8	11/6
5. Multicenter Long-term (40wk)	2.5	62	148/100	13/8	12/9
	5.0	71	145/101	14/10	13/9
	HCTZ 50	54	145/101	12/10	11/9

*Underlined values significantly different from placebo

Efficiency

It is clear that responder (and possible non-responder) subpopulations based on race exist for beta blockers, ACEI's, and AIIB's.

In addition to conducting initial dose finding studies in responder subgroups, could plan studies in broader population that explicitly planned to look for response differences based on

- Race
- Background therapy (include patients with and without diuretics and other drugs in same study). Most rigorous is the factorial dose-response study; develops data for single drug and combinations.
- Renin status (aldosterone levels, etc.)
- Genetic markers

Efficiency

Other areas (in addition to hypertension) where looking at responders in dose response studies could be useful:

- Asthma drugs (Cromolyn-anecdote)
- All symptomatic GI conditions (GERD)
- Anti-arrhythmics

Efficiency

c. Examine maintenance dose

When dose-finding occurs, it is almost always during initial treatment. For long half-life drugs particularly, but others too, examining the maintenance dose response could be very useful

- If 20 mg of fluoxetine works acutely and drug and metabolites have half lives well > 1 week, the maintenance dose is surely well under 20 mg, based solely on PK arguments. Lower maintenance doses could lead to a wide range of safety advantages. This has never been studied

Efficiency

- It would not be surprising if the dose needed to treat acute exacerbations of mania, depression, and other diseases was larger than the dose needed to maintain patients. Perhaps alternate day dosing would work. (Could this be true for Lotronex? Would lower doses have given less constipation or even less ischemic colitis?)

- Astemizole has a long half-life but was used acutely in seasonal allergies. It could have been used as a loading dose of 10 mg with subsequent lower doses $\leq$ 3 mg. That would have placed dose at about 1/5 of QT prolonging dose, instead of at 1/2. The drug might still be available

Efficiency

Maintenance dose-response studies are easy. Use randomized withdrawal design. People on treatment, doing well, are randomized to placebo and several doses of the drug

Conclusions

1. We've come a long way. We know how to get good D/R data and often get it

2. There are efficiencies possible that are, so far, rarely used

3. The biggest gap, perhaps highlighted by the promise of pharmacogenomically-desirable pharmacodynamic differences between individuals, is detection of individual D/R relationships. This will probably need new kinds of studies in which people are given more than one dose