Design of Clinical Trials

Robert Temple, M.D.
Deputy Center Director for Clinical Science
Center for Drug Evaluation and Research
U.S. Food and Drug Administration

FDA Clinical Investigator Training Course
November 13, 2018

Outline

1. Adequate and well-controlled studies and the 1962 FD&C Act
2. Non-inferiority studies
3. Enrichment designs

The Effectiveness Requirement

Until 1962 a drug had to be shown “safe for its intended use” to be marketed, but there was no requirement to show effectiveness. There was talk about effectiveness (how can a drug be safe if it has risks and provides no benefit) but no requirement and few studies we would recognize as well-designed and informative.

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 means. 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 requirement to show effectiveness 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 everything, all of medical science, really

- Such studies are the only basis for approval
- Note the plural. The agency interpreted this as requiring more than one controlled trial (modified by FDAMA 1997 to allow one study in some cases). Legislative history explicitly supports the view that congress meant more than one study.
- No relative efficacy. A drug need not be better than (or even as good as) other drugs (unless inferior effectiveness leads to lack of safety, this is Not in regulation but in 1995, an FR notice by Deputy Commissioner for Policy William Schultz. Noted that for treating a life threatening disease (stroke, MI or a contagious infections, it is critical that a new drug be as effective as approved drug.
- 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 the 1970's and 1980's 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 fully specified protocols and statistical plans, managing interim looks, maintaining blinding, avoiding loss of patients, dealing with 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) the study 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)

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.

[Note, our view on this has evolved; we would be very wary of a study whose methods of analysis were described only after unblinding]

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 much 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. A better term would be "blinded no-treatment control," as it would not imply that the change in the placebo group was a "placebo effect."

Ethics

Difference-showing

Blinded, randomized

No external data needed (assay sensitivity)

Baseline placebo is NOT a placebo concurrent control

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 will usually include a placebo control and may include an active control. Dose-comparison trials usually include randomization and blinding of patients or investigators, or both.

Effectiveness (positive slope or one

dose better than another) 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. Our initial recognition arose importantly from a historical error, the dose of diuretics

- Effective dose 1/8-1/4 of the dose recommended, studied, and used. (Seen in Materson 1978 study of chlorthalidone)
- Hypokalemia, almost surely decreased survival benefit of treatment by provoking arrhythmias. Also caused gout. The 100 mg chlorthalidone dose was stopped because of excess mortality in large NIH trial.
- In the 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. Could include a blinded endpoint adjudication committee.

- Many examples: GUSTO, GISSI, cancer trials
- Need objective endpoints but what is objective is not always so clear (ART, LRC)
- Other kinds of possible bias: other treatment, interpreting endpoints or referring endpoints for adjudication

Recent concern regarding the RECORD study of Avandia, (rosiglitazone) could referrals of cases for adjudication have been influenced by knowledge of treatment.

Adequate and Well-Controlled Studies (Cont'd)

(IV) Active Treatment Concurrent Control. The test drug is compared with known effective therapy; this design would be used, 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. This concern about "assay sensitivity" of the trial has been greatly elaborated in ICH E-10 and in FDA's non-inferiority guidance.

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. Usual use we see is to show similarity and conclude new drug is effective (like the control drug).

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 absent treatment).

Often used in oncology (tumors do not shrink) and for well-understood genetic diseases. You MUST have good, up-to-date natural history data.

Historical Controls

Critical Reference -

Sacks, Chalmers, Smith
Am J. Medicine (1982); 72:233-240.

Comparison of RCTs and HCTs for same disease

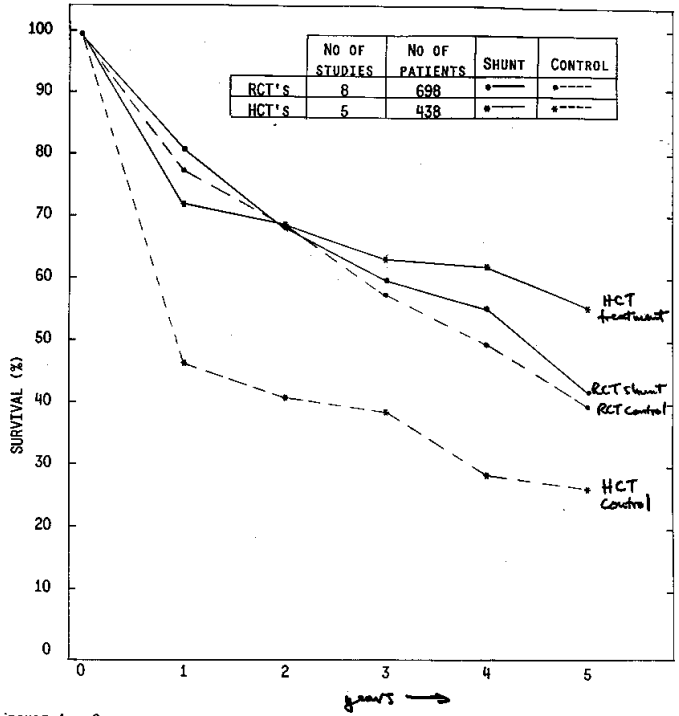
<u>Always</u>

1. RCT less favorable than HCT
2. Reason was that the historical no treatment control was worse than the randomized no treatment 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	
	Effective	Ineffective	Effective	Ineffective	Effective	Ineffective
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



IGURE 1. SURVIVAL OF TREATED AND CONTROL GROUPS IN CLINICAL TRIALS OF SHUNT URGERY 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 – an illustration of potential error

Gocke (letter to NEJM in 1970) observed 9 consecutive cases of acute fulminant hepatitis B, 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 as definitive but concluded it could represent better care, earlier Rx.

Therefore RCT in severe hepatitis B, comparing hyperimmune globulin to normal serum globulin. There was about 60% survival in both groups; i.e., no effect of treatment.

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 before 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 and 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) – can use ITT analysis
     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, perhaps with stratification for important baseline properties
   - Demography
   - Disease severity, risk factors
   - Other treatment, current or past
   - 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, e.g. exercise, breathing
    Exclusion of patients - ART
    Eligibility
Differential drop-outs – "informative censoring"
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 likelihood 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.

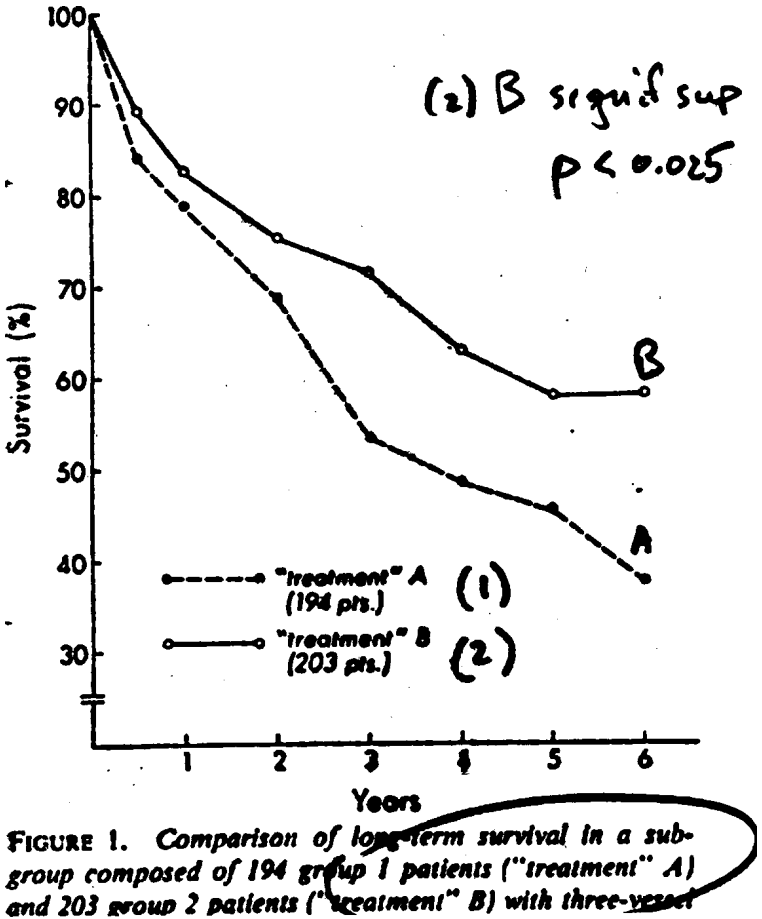


FIGURE 1. *Comparison of long-term survival in a subgroup composed of 194 group 1 patients ("treatment" A) and 203 group 2 patients ("treatment" B) with three-vessel*

Lee, et.al.

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

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.

Leaving aside lack of clear statement as to primary endpoint, the death and sudden death findings were all wrong because

1. Eight deaths in patients randomized to Anturane and one on placebo were dropped from the analysis because they were found "ineligible" (6) or poorly compliant (3).
2. Cause-specific mortality was unreliable. Different death causes, e.g., sudden death and AMI and "other" often had the same description; cases were called "sudden" death in placebo group and the same descriptions were called "MI" or "other" on Anturane, supporting an Anturane effect or "sudden death."

A.R.T. REPORTED MORTALITY RESULTS

	Pl	S	% ↓ (p)
PATIENTS (Eligible)	783	775	
ALL DEATHS (analyzable)	62	44	29% (p=0.076)
CARDIAC D's	62	43	30.6 ~~32%~~ (p=0.058)
SUDDEN	37	22	43% (p=0.041)
AMI	18	17	--
OTHER	7	4	--
OTHER CV	0	1	--

MORTALITY by CAUSE, TIME

	Pl	S	% ↓ (p-value)
ALL CARDIAC	62	43	30.6% (p=0.058)
ALL CARDIAC			
0-6 M	35	17	50% (p=0.021)
7-24 M	27	26	
SUDDEN			
0-6 M	24	6	74% p=0.003)
7-24 M	13	16	
NON-SUDDEN			
0-6 M	11	11	
7-24 M	14	10	

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 bias toward all-cause mortality, but a problem if the population has high non-study related mortality.

   - often accept CV mortality (but without trying to distinguish cause of death 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. We are somewhat more willing after a “win” on the planned primary endpoint.

3. Insist on full accounting of all randomized patients and a full on-treatment or ITT analysis (even if sponsor prefers another).removal of on-treatment deaths, as done in the ART would not be accepted.

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 or measure of function, such as exercise test
- 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/HbA1c, LDL cholesterol).

But experience with antiarrhythmics, inotropic drugs for heart failure, and more recently experience with erythropoietin to raise HCT and torcetrapib (raises HDL cholesterol) has led to considerable caution.

A rule (1992) on “Accelerated Approval” addressed this, reflecting both skepticism and the sense of urgency that can arise in relation to drugs for 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) oncologic drugs (response rate in refractory disease and time to progression), and for orphan drugs, where mechanism of action may be well understood.

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. A comparative study that showed no difference between symptomatic treatments would rarely be interpretable without a placebo group.

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 can lead to a falsely negative study, but does not lead to an erroneous conclusion that the drug is effective.

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 (known, really) to have in the study. This is the largest possible non-inferiority margin, M_1, and ruling out a difference as large as M shows the test drug has some effect. Usually, the margin is smaller than that, M_2, and is chosen to assure that the drug has a clinically meaningful effect.

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 that the inferiority of the new drug (C-T) is no greater than a specified margin M

$C-T<M$

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

[Old, naïve way (but still sometimes seen in 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

Placebo controlled trials have as a null hypothesis that the effect of the test drug (T) is ≤ 0 (placebo).

Ho: $T \leq P$

Ha: $T > P$

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, (rules out Ho, the null hypothesis) 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 some specified difference or margin (the non-inferiority margin M).

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).

But the effect of the control that determines M is not measured in the study and must be estimated/assumed based on the effect of C in previous studies.

M is Crucial

Everything depends on the validity of M; if the chosen M is larger than the actual effect of C in the study, e.g., if C had no effect in that study or an effect smaller than M, you will reach an erroneous conclusion that T is effective. If, e.g., you say M=10, then if C-T (97½% CI upper bound) is < 10, say 8, you would conclude that T has an effect. But if in the study the effect of C was in fact only 5, T would NOT have had an effect.

IT WOULD ONLY LOOK LIKE IT DOES

So 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 some 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 the trial shows that not all of the effect of C was lost (C-T < M_1), 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 in this study (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. If the trial did not have assay sensitivity, then even if C-T < M, you have learned nothing about the effect of T because the control did not have an effect on M.

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

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 looks as if some trials are different from others; i.e., there is a treatment by study interaction.

YOU CANNOT USE AN 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. Note that we generally consider effect on relative risk to be more stable than effect on absolute risk.

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, not included in the outcome trials of ACEI's in CHF.

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

Active Controls Equivalence Credible

1. Most serious 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), and more recently, many others
6. Anesthetic agents
7. Beta-agonists in bronchospasm
8. Comparison of anticoagulants in chronic AF
9. Comparison of anti-platelet drugs in ACS (Probably)

Four Critical Steps in Using a Non-Inferiority Design

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

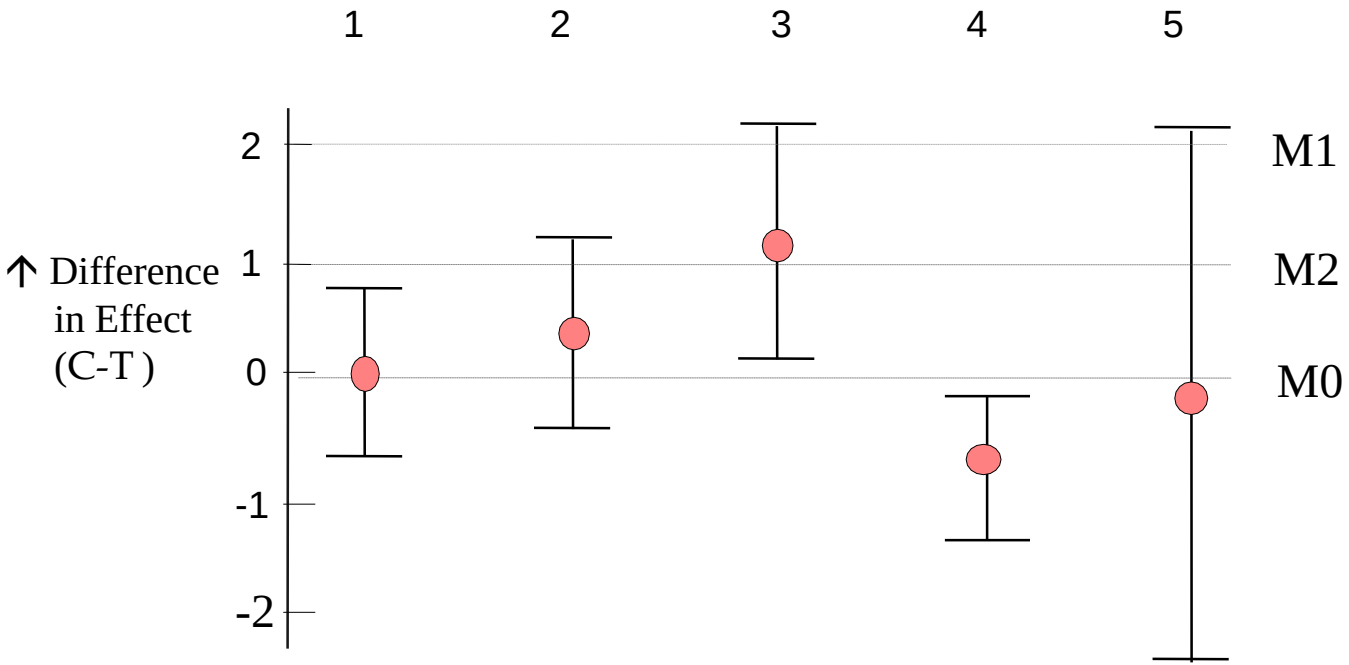
2. Setting an acceptable non-inferiority margin, M_1, a margin no larger than the effect the control can be reliably presumed to have had in the study, and that also reflects the fraction of the control effect that is considered clinically essential, M_2

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

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 some 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 (Prognostic Enrichment)

Although the information distinguishing individuals with respect to risk is growing exponentially, we've had such information before

- Epidemiologic risk factors for likelihood of cardiovascular outcomes
  - Severity of heart failure
  - Cholesterol, blood pressure levels; angiographic appearance
  - Diabetes
  - Recent events (AMI, stroke)
  - Elevated CRP (JUPITER Study of rosuvastatin)
  - Family history
  - Gender, race, age
- Risk factors in cancer
  - Previous breast cancer to predict contralateral tumor
  - Tumor histology or genetic/proteomic markers to predict occurrence or metastases

Prognostic Enrichment

1. Oncology

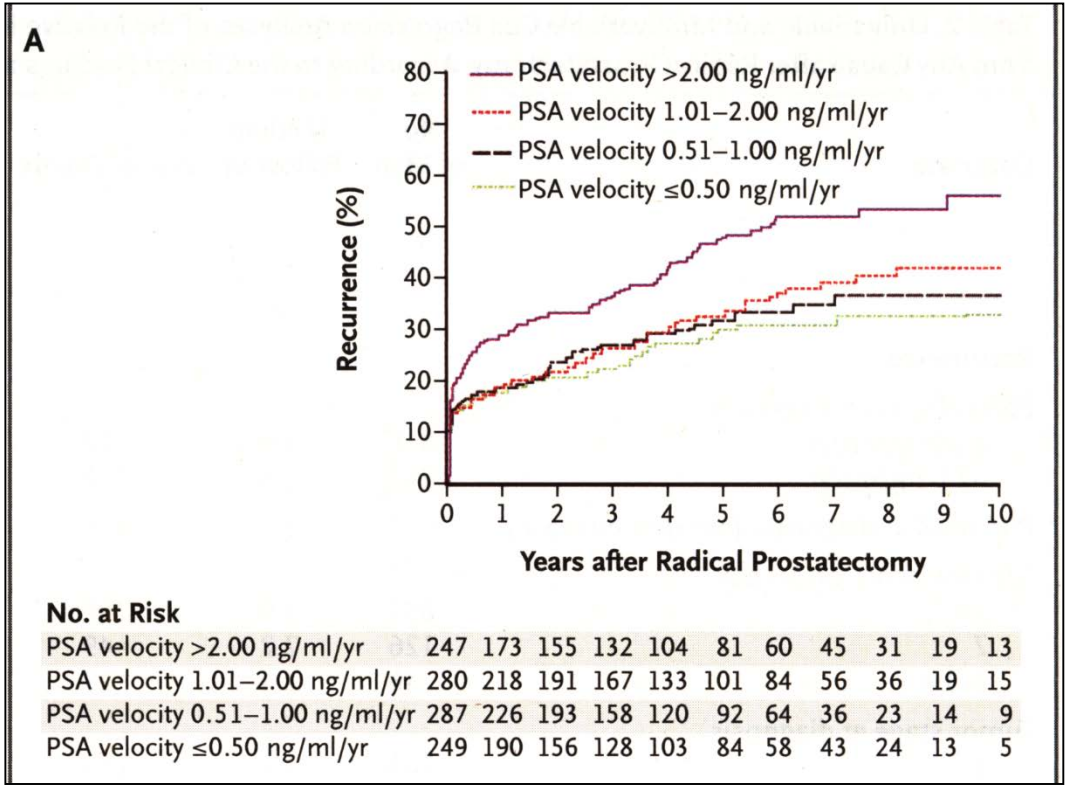
Prognostic enrichment would be critical in any study of chemoprevention or of any adjuvant chemotherapy. Tamoxifen prevented contralateral breast tumors in adjuvant setting (very high risk); it was then studied in people with more general high risk. This was needed a) to have enough endpoints to detect 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 of tamoxifen (% reduction) in the people selected, but more events would be prevented.

Prognostic Enrichment

1. Oncology (cont.)

   Potential (not used or maybe not fully accepted, but a good illustration) selection method for patients with more frequent endpoints in prostate cancer adjuvant treatment:

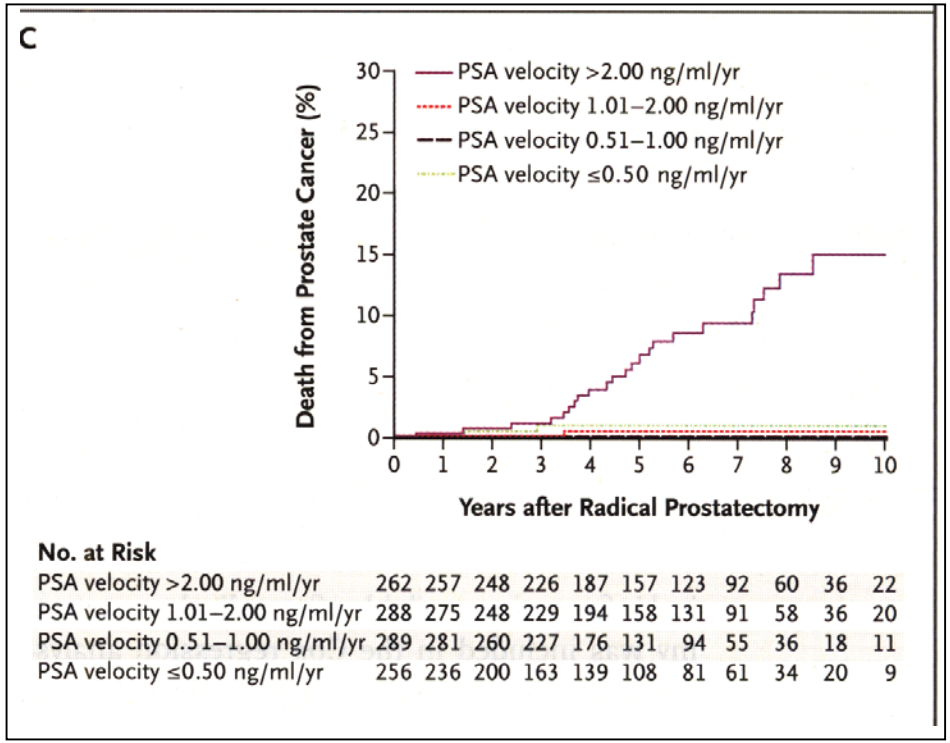
   D'Amico reported [NEJM 2004; 351:125-135] that in men with localized prostate Ca, following radical prostatectomy, PSA "velocity" (PSA increase > 2 ng/ml during prior year) predicted prostate Ca mortality almost 100% over a 10 year period. There were essentially no deaths from prostate Ca (many from other causes), even though recurrence rates were not so different. Given concerns about effects of treatment on survival, an adjuvant prostate Ca study would surely want to include patients at risk of death.



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



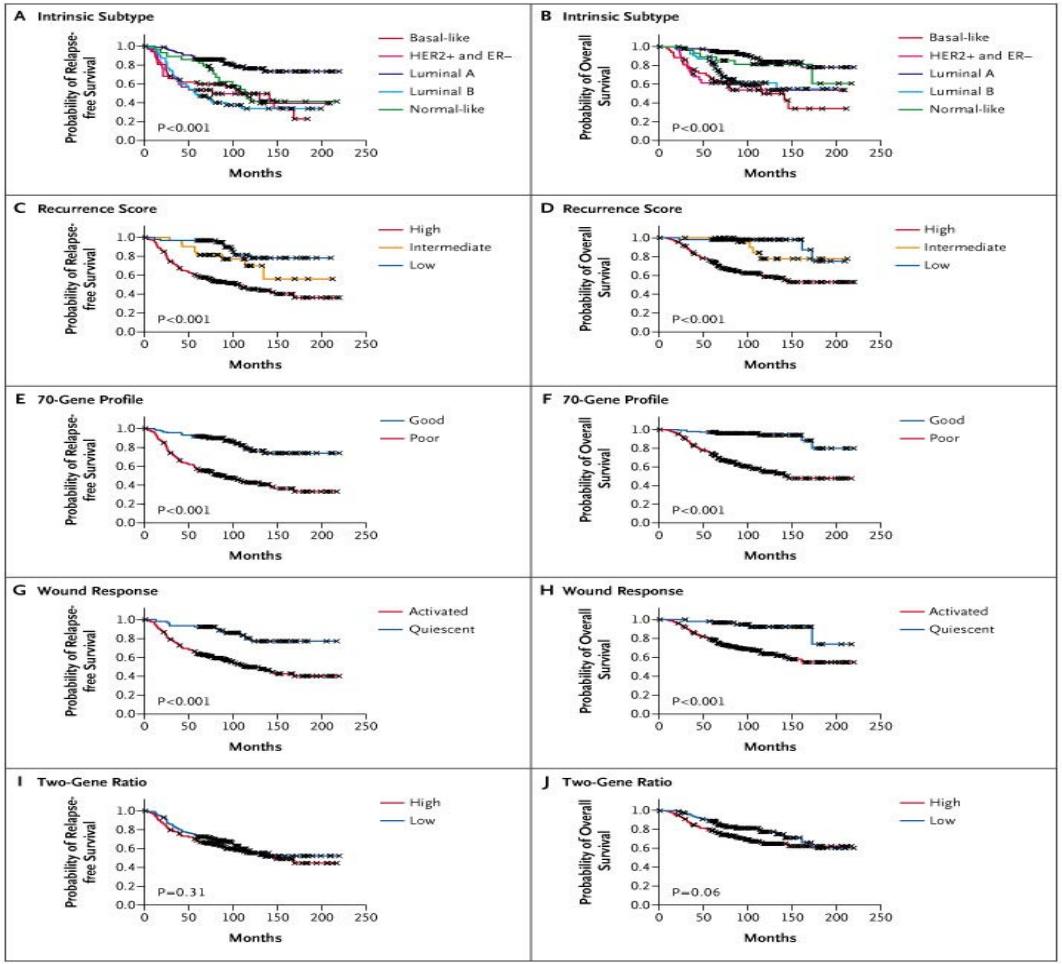
Kaplan-Meier Estimates of the Cumulative Incidence of Death from Prostate Cancer (Panel C) after Radical Prostatectomy, According to the Quartile of PSA Velocity during the Year before Diagnosis

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 market test called Mamaprint is now available for this assessment.


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

Prognostic Enrichment

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) was an outcome study in NYHA class IV patients. It included 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) was in a post-MI, very high cholesterol population: 9% 5 year CV mortality. It showed a survival advantage much harder to show in later studies.

- JUPITER study of rosuvastatin included people with "normal" LDL but high CRP.

Prognostic Enrichment

3. Other

Identifying people at high risk is especially important in “prevention” or risk reduction efforts, as the CV and oncology examples indicate. There are many other areas where this would be important, notably for preventing or delaying the development of Alzheimer’s Disease, where it may be necessary to treat before there are manifestations of dementia. It has been suggested that people with minimal brain dysfunction or other early abnormalities might be suitable. A population without such a predictor might have few or no cases over many years, making a demonstration of an effect impossible.

Predictive Enrichment

Probably the most exciting enrichment strategy today is predictive enrichment, finding the patients with the greatest likelihood of responding to treatment. This represents the “individualization” of treatment we all dream about. Studying people who will respond to a treatment greatly enhances the power of a study, facilitating approval, but it may also have critical implications for how a drug will be used.

It can be especially important 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.

Selection can be based on understanding of the disease (pathophysiology, tumor receptors) or it can be empiric (e.g., based on history, early response.

There are many examples in oncology related to proteomic or genomic responses. This is perhaps not surprising as cancer is a “genetic disease.” I will also consider more “empiric” examples where we may not understand the predictive markers. More recently, genetic subtypes of cystic fibrosis and hepatitis C have been shown to respond dramatically to new treatments; where these are low in frequency a study in the overall population with this disease would probably have failed.

Predictive Enrichment

Pathophysiology

- 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; or, rather, we analyze the patients who turn out, after randomization, to have a sensitive organism.

- A well-established genetically determined difference could be the basis for a pathophysiologically selected population. Many tumor genetic or surface markers are related to well-understood effects on enzymes or tumor growth rates; Herceptin for Her2+ breast tumors; selection of ER^+ breast tumors for anti-estrogen treatment, and use of many other receptor markers illustrate this.

Predictive Enrichment

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

- CAST was carried out in people who had to have a 70% reduction of VPB's during a screening period. Only "responders" were randomized. Trial showed harm, not benefit, but properly tested the question, as previous trials had not.
- Beta-blocker CHF trials were carried out only in people who could tolerate the drugs.
- 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, often to a dose-response study (note, by the way, that one could argue that all D/R studies should be done in responders, including non-responders flattens the D/R curve).
- Every randomized withdrawal study has this characteristic (more later).

Predictive Enrichment

As noted, (CAST, Oates) selection could be based on response of a biomarker; that is, screen the entire group and randomize only those with a good response.

Other 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) – never tried, to my knowledge
- Only patients with CRP response > x
- Only people who make the relevant active metabolite (clopidogrel)

Advantages of Predictive Enrichment

1. Efficiency/feasibility

When responders are a small fraction of the population, predictive enrichment can be critical.

Table 2: Sample Size Ratios as a Function of the Prevalence of Marker-Positive Patients

Prevalence of Marker-Positive Patients	Response in Marker-negative Patients (% of marker positive response)	
	0%	50%
	Sample Size Ratio	Sample Size Ratio
100%	1.0	1.0
75%	1.8	1.3
50%	4	1.8
25%	16	2.6

Advantages of Predictive Enrichment (cont)

As the table shows, if 25% of patients have the marker that predicts effect and marker negative patients have no response, an unselected population would need 16 times as many patients [the gain is much less if marker negative patients have same response, even if it is smaller]. Recently, FDA approved ivacaftor for CF patients with a specific gene mutation that is present in just 4% of CF patients. A study in an unselected population would have had no chance of success. Similarly, boceprivir and telaprivir were shown to be strikingly effective in patients with type 1 hepatitis C virus, the type most resistant to standard therapy.

2. Enhanced B/R if there is toxicity (Herceptin).

Trastuzumab (Herceptin) is cardiotoxic. Studies in patients with metastatic cancer as well as adjuvant studies were conducted in patients with Her-2-neu positive tumors, enhancing B/R. Her-2-neu negative patients have much less response, and the cardiotoxicity is unacceptable.

Data in the Marker-Negative (Off) Group

Two important questions arise when using such selection criteria. One is the quality of the genetic or other predictive test. The second is the sensitivity and specificity of the various predictive cut-off points (how positive must Her-2-neu be?) In general, unless there is no real chance of an effect in marker-negative patients, some negative patients should be included in studies (stratified) because

- They may have some response
- They may help refine the marker cut off

Early studies may solve this problem, but the larger numbers in later trials may give better answers. It would still be possible to make the primary endpoint the effect in the enriched stratum (routine in antibiotic trials where sensitivity of the organism is not known at randomization), while examining response in patients below the cut-off .

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.

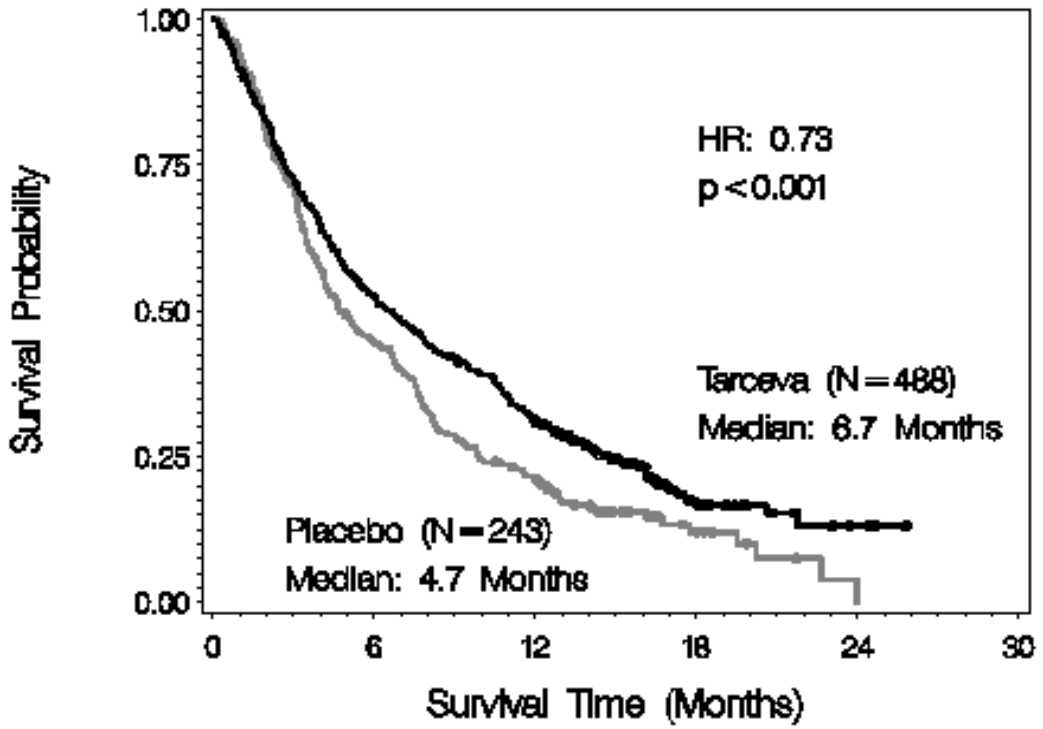
Selection of Likely Responders

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%		

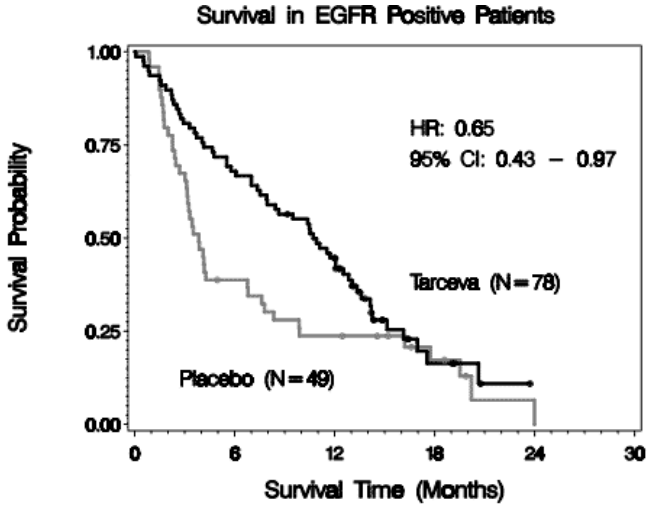
Kaplan-Meier Curve for Overall Survival of Patients by Treatment Group

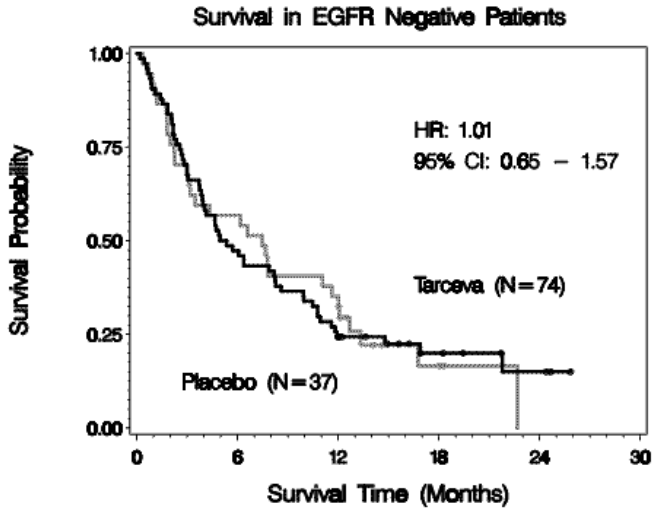


Tarceva (erlotinib)

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)

Predictive Enrichment – Pathophysiology or Genetic Characteristics

1. Only people who make the active metabolite (clopidogrel)
2. Only people whose tumor takes up the drug (History, test for I 131 uptake in thyroid tumor to choose dose)
3. Effect on tumor metabolism, e.g., glucose uptake
4. Proteomic markers or genetic markers that predict response

Plainly, the wave of the future in oncology (Herceptin; imatinib inhibits c-KIT, a receptor for tyrosine kinase, that is mutated and activated in most GIST patients; vemurafenib in melanoma effective in patients with activating mutation $BRAF^{V600\text{-}E}$.

Usually the marker is pre-selected but Friedlin and Simon suggest a way to look for responsive subsets half-way and analyze both whole population and subset.

Predictive Enrichment - Adaptive

1. Simon proposal

Rich Simon has suggested a design potentially useful where you do not have an identified predictive marker.

1. Design study as usual, but divide into first half, second half.

2. Run first half of study and search for genetic predictor of response (any analyses, as many as you want)

3. Complete the study, entering all patients (responders predicted and not predicted) but stratifying them

4. Divide study alpha as 0.04 for whole study and 0.01 for the response-predicted subset in 2nd half.

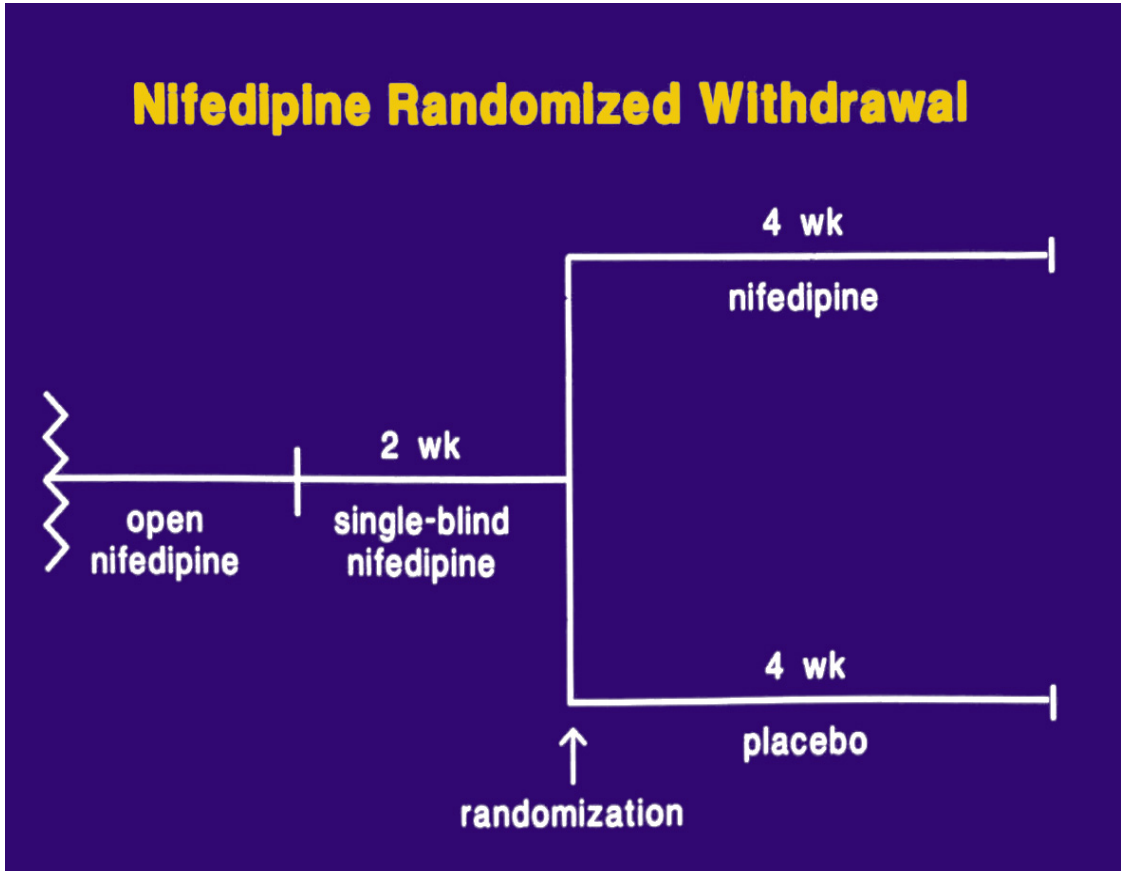
Randomized Withdrawal

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; apparent responders are then 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). Now standard for maintenance studies in depression and psychosis. NEJM, reported a trial that compared rates of, recurrence of psychosis or agitation in patients with Alzheimer’s Disease who had responded to risperidone for 16 weeks and were then randomized to placebo as confirmed respondent

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

Early use in studying nifedipine in vasospastic angina (first approved use) after advisory committee rejected a baseline controlled study. Note small study (n = 28) and lack of recurrence in 9/15 on placebo.


Nifedipine Randomized Withdrawal
4 wk
nifedipine
2 wk
open
nifedipine
single-blind
nifedipine
4 wk
placebo
randomization

	Nifedipine	Placebo
n	13	15
Early withdrawal	0	5*
Early withdrawal plus AMI	0	6*
Investigator's judgment of success	11	2*
Median angina/week	0	3.4*
Mean angina/week	0.7	18.4*
Change from baseline in attacks/week		
better (≤ 1)	0	0
same (± 1)	11	6
worse (≥ 1)	2	9

*$p<0.05$, one sided

Randomized Withdrawal

The randomized withdrawal study can also be 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 BP effect in hypertension (long-term placebo would be unethical)

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

Randomized WD – Another Possibility

There is growing concern about how to analyze drop-outs in clinical studies and recent NAS report identified “not having them” as the best method. In symptom trials, however, where we want evidence of persisting effect (e.g., in pain studies), drop-outs are hard to avoid.

A possible approach in these cases is to use short (4 week) studies as initial evidence of effect, followed by a trials in which known (apparent) responders are followed for, say, 12 weeks, after which they enter a randomized WD study of short duration, e.g., 2 weeks or until pain returns. There would be few dropouts in the WD study and, in some sense, it asks the pertinent question:

In patients who respond, does the effect persist (it can’t

persist in the non-responders).

There designs are being used in narcotic pain trials.

Randomized Withdrawal (cont.)

Design has major advantages

- Efficient: “enriched” with responders, so will show a larger drug-placebo difference
- Efficient: patients already exist and known, e.g., a part of an open or access protocol
- Ethical: can stop as soon as failure criterion met, very attractive in pediatrics

Other Predictive Enrichment

Studies in non-responders; randomize to new drug and failed drug. A comparison enriched with people who will not respond to the control drug, increasing drug-control difference.

Studies in intolerants; randomize to new drug and poorly tolerated drug, a comparison enriched with people who will do “badly” on the control drug.

Both should give a larger drug-control difference.

Very valuable findings – rarely attempted.

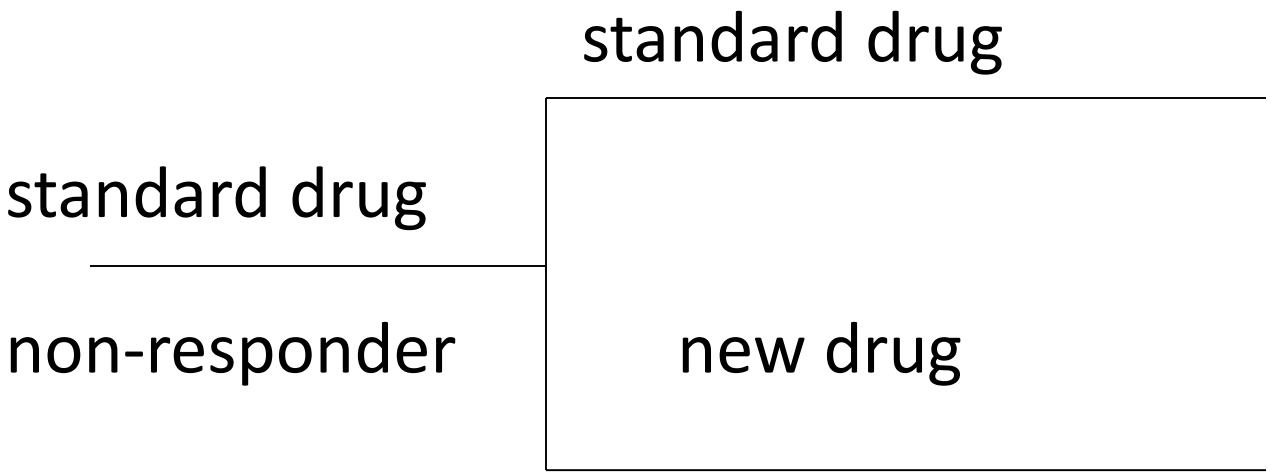
Studies in Non-Responders

Design should give the new drug an edge (they've failed the other) and it has allowed approval of drugs otherwise too toxic

- Captopril (thought to cause agranulocytosis) was superior to diuretic, reserpine, hydralazine (triple therapy) in patients failing triple therapy.
- Bepridil (a CCB) superior to diltiazem for angina in diltiazem failures.
- Clozapine superior to thorazine in standard therapy failures.

The design must randomize to failed and new drug.

Studies in Non-Responders



Clozapine

Too toxic unless clear clinical advantage

Study in schizophrenics unresponsive to standard therapy

- History of poor response to neuroleptics
- Diagnosis of schizophrenia, hospitalized
- 6 week failure on haloperidol
- 4 week, double-blind comparison of clozapine vs. chlorpromazine plus benztropine

Results

	Response (%)	
	Clozapine	**CP2**
CGI (decrease ≥ 1)	71	37*
BPRS items (dec ≥ 1)		
concept disorganization	60	39*
suspiciousness	64	42*
hallucinations	59	51
thought content	65	40*
CGI and BPRS	15	2*

*p ≤ 0.05

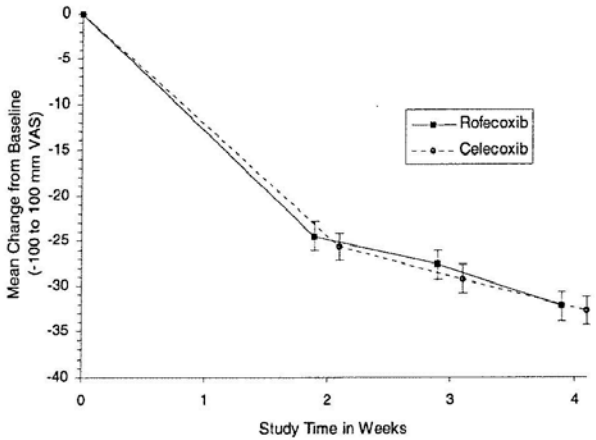
Studies in NRs

It does not always work, though. In discussions of NSAIDs, all arthritis doctors said many drugs are needed because responses are individual. Plausible, but at a COX2 meeting a few years ago I suggested studies in NRs.

Merck did a study comparing rofecoxib 25 mg and celecoxib 200 mg in celecoxib non-responders.

Figure 1

Pain at Night while in Bed (WOMAC)
Mean Change from Baseline ± SE by Timepoint
(Primary Per-Protocol Approach)



Note that without a celecoxib control, rofecoxib would have appeared VERY effective in this NR population.


FDA
U.S. FOOD & DRUG
ADMINISTRATION