Evolution of Clinical Trial Designs

Robert Temple, MD
Deputy Center Director for Clinical Science
Center for Drug Evaluation and Research
Food and Drug Administration

American Society for Quality
February 12, 2015
Overview

I will first talk about how the basic requirement for approval of drugs, evidence of safety, and substantial evidence of effectiveness based on adequate and well-controlled studies, came to be developed including some of what we learned as we implemented the law, regulations, and our extensive guidance.

People learning and practicing medicine today cannot really imagine what the basis of medicine was like even in the 1950’s and 1960’s. Apart from obvious things (removing an appendix, curing infections, Lasix), it was hard to say what we really knew:

- Controlled trials were hardly ever seen – effectiveness was rarely convincingly established
- Drug labeling was fantasy (Early 1960’s PDR listed 50 treatments for alcoholism)
- Outcome trials basically didn’t exist till VA studies in HT (1967), UGDP (late 1950’s)
- Statistical inference was primitive once you left a few special places (NIH, mostly, and perhaps analgesia)

And then it all changed, beginning in 1962
The Ages of Drug Development and Drug Regulation

Age of Safety - 1938
Age of Effectiveness - 1962
Age of Individualization and Dose-Response
- from early 1980’s

AGES CUMULATE AND CONTINUE TO DEVELOP AS WE LEARN
The Ages of Drug Development and Drug Regulations

Pre-History (pre-1938)

FDA existed (1906) but could only respond to problems. There was complete freedom to market; no requirement for testing or approval; Government could seek to remove dangerous or misbranded products. There were some disasters:

DNP - weight loss drug, caused thousands of cataracts, enucleations in 1930’s

1937 - Elixir Sulfanilamide; killed over 100 people, many of them children; diethylene glycol (anti-freeze) was used to dissolve the drug. There were no animal tests; the chemist smelled and tasted it. This episode led to:

The Age of Safety; the Federal Food, Drug and Cosmetic Act of 1938
The Food, Drug and Cosmetic Act of 1938

Required:

1. Pre-market notification. Marketing required an approved new drug application (NDA), but the NDA became “effective” (approved) in 6 months if FDA did not object; time could be extended. Reflected a strong expectation that there would be approval.
The Food, Drug and Cosmetic Act of 1938

2. Required a demonstration of safety. The application could be refused if

(a) Investigations did not include all tests reasonably applicable to show whether drug is safe when used under proposed labeling

(b) Results of tests show unsafe or do not show that it is safe

(c) Information submitted or any other information available are insufficient to determine whether safe

(d) Labeling is false or misleading in any particular

These safety requirements of 1938 are identical to current requirements. Note how broad and possibly subjective they are. What does “safe” mean? That benefits outweigh risks. But there was no good evidence of benefit.
The Age of Safety

We weren’t very good at safety evaluation in the early days

No standards
No controlled trials
No post-marketing surveillance

Result: Some very toxic drugs were approved and toxicity remained unrecognized, e.g., iproniazid, isoniazid.

Iproniazid (Marsilid) and Isoniazid (INH), both approved in about 1950, were as hepatotoxic as any drugs ever marketed, but

• the first reports of liver injury for iproniazid took 3 years and it was not withdrawn until 1956

• INH was thought mildly hepatotoxic until a PHS trial in young men in 1971 showed mortality of 0.1%

In contrast, we now catch hepatotoxins pre-marketing (often – dilevalol, tasosartan) and very quickly if not (bromfenac, troglitizone within months)
The Age of Effectiveness - 1962

Why thalidomide (a safety problem) led to a requirement for demonstrating effectiveness is not obvious, but it put the FD&C Act “in play” and then anything can happen. And these had been several years of discussion before Sen Estes Kefauver by clinical pharmacologists on the terrible data available. Actually, the 1962 Act made at least 3 important changes:

1. FDA had to give positive approval before a drug could be marketed

2. A meaningful requirement to study drugs under an IND and an explicit requirement for informed consent (one year before the Declaration of Helsinki).

3. The effectiveness requirement

It also required review of all the drugs approved 1938-62 on the basis of “safety” to determine their effectiveness. A huge, but successful, effort started by the NAS/NRC, leading to withdrawal of about one third of existing drugs and many claims for the ones that remained. The DESI project.
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 itself that was radical. I believe we might have imposed that by regulation, as “safe for intended use” implies a risk/benefit analysis, i.e., need for evidence of benefit. It was the need for adequate and well-controlled studies, which were the only basis for concluding a drug was effective, that changed everything, all of medical science, really

- These are the only basis for approval
- Note the plural. The agency interpreted this, with support from legislative history, as requiring more than one 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

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 it was a very clever and purposeful trade-off. “Substantial,” legally, is a low standard (between a scintilla and a preponderance)

But adding the need for A&WC studies as the only source of substantial evidence (indeed, two of them) turns a low standard into quite a high one [especially with the p<0.05 (two-sided) statistical significance standard 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 learned. I will illustrate some of what we learned later, including

1. Counting all patients and risks of assessing cause-specific events – the Anturane Reinfarction Trial
2. Interpretation of active control non-inferiority trials
3. Good dose response designs

There are MANY other critical aspects of studies including specifying planned statistical analyses use of surrogate endpoints, accounting for multiple endpoints and, of course, how to assess safety.
No law or disaster led to this third age. In the late 1970’s, we began to understand how to assess dose-response, ultimately leading to 1993 ICH E4: “Dose-Response Information to Support Drug Registration”

Serious drug-drug interaction experience (terfenadine), that is where one drug affects the blood level of another drug, has led to a revolution in assessing metabolism of drugs and how they interfere with the metabolism of other drugs or are themselves affected by other drugs.
Age of Individualization

From the early 1980’s, there was growing interest in effects in demographic subsets of the population (age, gender, race) and in populations defined by excretory abnormalities, generally, loss of kidney function. We have written guidance on inclusion of both genders (1993) and the elderly (1989) in studies, and have encouraged and required pediatric studies (often post-marketing).

By regulation in 1998, NDA applicants must analyze safety and effectiveness data by age, gender, and race.
Age of Individualization

We are in the middle of a major refinement, the result of pharmacogenomic information. Most individualization up to now was based on PK differences, because they are very easy to measure. E.g., you lower the dose for small people or people with reduced kidney function.

What about PD differences; i.e., a group that responds more (or less) to a drug or is more (or less) susceptible to adverse effects? We say and believe that individuals respond differently to NSAIDs, SSRI’s, various asthma drugs, but that is hard to show in studies.

Pharmacogenomics will alter all that. We are finding that people with particular tumors, particular genetic causes of diseases like cystic fibrosis, respond to specific drugs. We’re at the beginning.
Legal Requirements
Food, Drug and Cosmetic Act
More Detail

Sec 505

(a) Need an approved application to market

(b)(1) Application has full reports of investigations to show whether drug is safe and effective and detailed components, composition, methods and controls

(c) FDA must give positive approval (changed in 1962 from 1938 - then NDA became effective if FDA did not respond)
Legal Requirements (Cont’d)

(d) Grounds for Refusal to Approve

(1) Investigations do not include all tests reasonably applicable to show whether drug is safe and used under proposed labeling.

(2) Results of tests show unsafe or do not show that it is safe.

(3) Manufacturing methods and controls are not adequate to preserve identity, strength, quality and purity.

(4) Information submitted or any other information available are insufficient to determine whether safe.

ALL UNCHANGED SINCE 1938
Legal Requirements (Cont’d)

(5) 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 (What) – new in 1962

(6) Labeling is false or misleading in any particular

(7) Section 502(f): A drug is misbranded unless its labeling bears adequate directions for use
Legal Requirements (Cont’d)

Only the Substantial Evidence was new in 1962; rest unchanged from 1938

The **only** basis for approval is data from adequate and well-controlled studies (plural, but modified in 1997).

Must provide all relevant information in full details. **NOT SUMMARIES**

No relative efficacy: a new drug need not be better than, or as good as, another (unless safety).

Effect must be clinically meaningful (court)

*SAFE* for intended use implies risk/benefit
Regulations

The clinical parts of an application are affected mainly by 3 regulations, the first 2 revised in 1985, the third created in 1992

21 CFR 314.50: Content and Format of an application

21 CFR 314.126: Adequate and Well-controlled studies

21 CFR 314.500: Accelerated approval (use of surrogate endpoints and approval with restrictions)
Clinical Section

Rule calls for

1. Description and analysis of every clinical pharmacology study and every controlled study, including the protocol and statistical analysis, as well as sufficient reports of everything else that is pertinent to safety and effectiveness from any source.
Clinical Section

2. Integrated summary of data showing substantial evidence of effectiveness and evidence to support dosage and administration, modifications for subgroups (pediatrics, geriatric, renal failure)

3. Summary and updates (4 months prior to approval) of safety information with all available information related to safety, including animal data, adverse effects, drug-drug interactions
Clinical Section

4. Case report forms for each patient who died or did not complete study because of an adverse event (thought drug related or not). Others on request. Prior to 1985, all CRFs required

5. Case report tabulations (replaced “all CRFs”)

   - All data from well-controlled studies
   - All data from earliest clinical pharmacology studies
   - Safety data from other studies

   Original intent was archival, but we have moved far toward usable data sets
Adequate and Well-Controlled Studies

314.126 Adequate and Well-Controlled Studies

Very critical, so I will review in detail, with discussion of the most important issues.

Only basis for approval

Apart from acceptable design and analysis (A and WC) the studies must show effectiveness, ordinarily a statistically significant effect on a meaningful endpoint, usually replicated.
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, a group very similar to the test group, and treated the same as people getting the test drug, except for getting the drug, 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, i.e. use a historical control.)
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
§314.126 Adequate and Well-Controlled Studies

The purpose of conducting clinical investigations of a drug is to distinguish the effect of a drug from other influences, such as spontaneous change in the course of the disease, placebo effect, or biased observation.

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)

(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 (really, blinded no treatment 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

Difference showing trials

- Placebo control
- No treatment
- Dose-response (duration-response)
- Some active control
- Most historical control

Equivalence-showing (or really, non-inferiority showing) trials

- Most active control
- Some historical control
Adequate and Well-Controlled Studies (Cont’d)

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.
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 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 that are different for the 2 groups
   avoidable (bias, unblinding) – population choice in historical controls

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
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.

Ethics
Difference-showing
Blinded, randomized
No external data needed for interpretation if positive (e.g., assay sensitivity)

But even a placebo-controlled trail has to be analyzed properly, as we learned form the Anturane Reinfarction Trial (ART).
Counting All Patients

In 1980 the need to account for all patients in a trial had never been an issue. Then came the Anturane Reinfarction Trial, a study, supported in the NEJM by two Dr. Braunwald editorials, that seemed to show a survival benefit in post-AMI patients treated with sulfinpyrazone (Anturane), an anti-platelet drug. Our analysis taught us a lot: about cause-specific mortality, multiple endpoints, unplanned 6 month subset analyses and complete follow-up [Temple R, Pledger G. The FDA's Critique of the Anturane Reinfarction Trial. N Engl J Med 303:1488-1492, 1980]
The Anturane Reinfarction Trial seemed a model effort, one of the first industry-sponsored outcome trials.
Features of A.R.T.

Double-Blind (U.A. values hidden) - Shipped from C-G with numbers.

Randomized in blocks of 10 within each clinic

Placebo-Controlled

Patient Population
- Male or female
- Age 45-70
- AMI 25-35 days before
- ECG Documentation
- Typical Pain History

Enzymes: 2 of CPK, SGOT, LDH had to exceed 2X normal - 72 hr

No cardiomegaly, CHF
>NYHA II, life-limiting disease

Baseline co-variates
- Index MI and later symptoms
- Smoking
- Medications
- Chest x-ray
# A.R.T. REPORTED MORTALITY RESULTS

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>S</th>
<th>%↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(p)</td>
</tr>
<tr>
<td>PATIENTS (Eligible)</td>
<td>783</td>
<td>775</td>
<td></td>
</tr>
<tr>
<td>ALL DEATHS (analyzable)</td>
<td>62</td>
<td>44</td>
<td>29% (p=0.076)</td>
</tr>
<tr>
<td>CARDIAC D's</td>
<td>62</td>
<td>43</td>
<td>30.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32% (p=0.058)</td>
</tr>
<tr>
<td>SUDDEN</td>
<td>37</td>
<td>22</td>
<td>43% (p=0.041)</td>
</tr>
<tr>
<td>AMI</td>
<td>18</td>
<td>17</td>
<td>--</td>
</tr>
<tr>
<td>OTHER</td>
<td>7</td>
<td>4</td>
<td>--</td>
</tr>
<tr>
<td>OTHER CV</td>
<td>0</td>
<td>1</td>
<td>--</td>
</tr>
</tbody>
</table>
## MORTALITY by CAUSE, TIME

<table>
<thead>
<tr>
<th></th>
<th>Pl</th>
<th>S</th>
<th>% ↓ (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALL CARDIAC</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL CARDIAC</td>
<td>62</td>
<td>43</td>
<td>30.6% (p=0.058)</td>
</tr>
<tr>
<td>0-6 M</td>
<td>35</td>
<td>17</td>
<td>50% (p=0.021)</td>
</tr>
<tr>
<td>7-24 M</td>
<td>27</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td><strong>SUDDEN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-6 M</td>
<td>24</td>
<td>6</td>
<td>74% (p=0.003)</td>
</tr>
<tr>
<td>7-24 M</td>
<td>13</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td><strong>NON-SUDDEN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-6 M</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>7-24 M</td>
<td>14</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
Ineligible Patients

It was not possible to see this from published reports, but 9 patients who had died were excluded from the results (8 Anturane, one placebo) for being “ineligible” or poor compliance (pills found in their room). When you put back exclusions, there was no documented effect.
## TOTAL CARDIAC DEATHS

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.R.T.</td>
<td>62</td>
<td>43</td>
</tr>
<tr>
<td>POOR COMPLIANCE</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>LATE INELIGIBLE</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>LESS THAN 7 DAYS</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>INELIGIBLE &lt;7D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>69</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>-0.2</td>
<td></td>
</tr>
</tbody>
</table>

## LATE DEATHS

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOTAL</strong></td>
<td>82</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>0.162</td>
<td></td>
</tr>
</tbody>
</table>
FDA guidance now clear in calling for full patient accounting. CONSORT statement similar. There is no doubt the 6 missing ineligible deaths had not been reported in NEJM.

We’ve also learned to be careful about cause-specific CV mortality. Cause specific mortality concern triggered by change in one patient cause from interim to final, triggering a “You mean these can change” thought and review of cases. We found that similar events were reported as SD on placebo, something else on Anturane, leading to a “reduced SD” claim. Here too we have been very cautious about such analyses.

We presume much of what happened occurred because the analysts of data were unblinded and could make these changes, leading to more attention to data access.
Another kind of controlled trial is the 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. 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. We have written much more about this in guidance.
Non-Inferiority Trials

Once called “equivalence trials” these are a major regulatory, ethical, international problem.

Fundamental distinction is between trials intended to show a difference between treatments and trials intended to show similarity or non-inferiority; latter pose major problems of interpretation.

Desire to use such trials is understandable: it seems sensible to compare new and old effective therapy, see no less of effect and declare victory. Avoids exposure of patients to ineffective treatment.
Interpretation of Active Control Trials

Active control equivalence or non-inferiority trials are the intuitive alternative to the placebo-controlled trial.

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. 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
The non-inferiority trial tries to prove effectiveness by showing that the difference between the new drug (T) and the control (C), i.e., C-T, is less than some margin (M), which is the effect of the control you know C had in this study. But M is not measured (there’s no placebo) so it must be assumed, based on past placebo-controlled trial experience. If you show statistically that

\[ C-T < M \ (97^{1/2}\% \ CI \ lower \ bound) \]

Then T has some effect > 0.
Interpretation of Active Control Trials (cont.)

The critical question is whether this trial could have distinguished the control from placebo. If it could have, the trial is said to have “assay sensitivity.”
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. If a trial has assay sensitivity then if $C-T < M$, $T$ had an effect. If the trial did not have assay sensitivity, then even if $C-T < M$, you have learned nothing.

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
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, because trials commonly fail to show drug is better than placebo. Trials of nomifensine illustrate this.
Assuring Assay Sensitivity in Non-Inferiority Trials – The Major Problem

We have found that about 50% of depression trials cannot distinguish effective drugs from placebo.

In almost all cases we therefore need a placebo-controlled trial in symptomatic conditions (or show superiority).

In outcome trials in contrast (antibiotics in serious pneumonia, UTI; trials of anticoagulants in AF) it has been possible to define non-inferiority margin.
TABLE 1. Results (4 week adjusted endpoint Ham-D total scores) of 6 trials comparing a new antidepressant, imipramine, and placebo showing only the new drug vs. imipramine comparison.

<table>
<thead>
<tr>
<th>Study</th>
<th>Item</th>
<th>Common Baseline</th>
<th>NEW</th>
<th>IMI</th>
<th>“p” two tail</th>
<th>Power to detect 30% difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>R301</td>
<td>HAM-D (n)</td>
<td>23.9</td>
<td>13.4</td>
<td>12.8</td>
<td>0.78</td>
<td>0.40</td>
</tr>
<tr>
<td>G305</td>
<td>HAM-D (n)</td>
<td>26.0</td>
<td>13.0</td>
<td>13.4</td>
<td>0.86</td>
<td>0.45</td>
</tr>
<tr>
<td>C311(1)</td>
<td>HAM-D (n)</td>
<td>28.1</td>
<td>19.4</td>
<td>20.3</td>
<td>0.81</td>
<td>0.18</td>
</tr>
<tr>
<td>V311(2)</td>
<td>HAM-D (n)</td>
<td>29.6</td>
<td>7.3</td>
<td>9.5</td>
<td>0.63</td>
<td>0.09</td>
</tr>
<tr>
<td>F313</td>
<td>HAM-D (n)</td>
<td>37.6</td>
<td>21.9</td>
<td>21.9</td>
<td>1.0</td>
<td>0.26</td>
</tr>
<tr>
<td>K317</td>
<td>HAM-D (n)</td>
<td>26.1</td>
<td>11.2</td>
<td>10.8</td>
<td>0.85</td>
<td>0.33</td>
</tr>
</tbody>
</table>
TABLE 2. Results (4 week adjusted endpoint Ham-D total scores) of 6 trials comparing a new antidepressant, imipramine, and placebo showing all comparisons.

<table>
<thead>
<tr>
<th>Study</th>
<th>Item</th>
<th>NEW</th>
<th>IMI</th>
<th>PBO</th>
<th>Baseline HAM-D adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>R301</td>
<td>HAM-D (n)</td>
<td>13.4</td>
<td>12.8</td>
<td>14.8</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33</td>
<td>33</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>G305</td>
<td>HAM-D (n)</td>
<td>13.0</td>
<td>13.4</td>
<td>13.9</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39</td>
<td>30</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>C311(1)</td>
<td>HAM-D (n)</td>
<td>19.4</td>
<td>20.3</td>
<td>18.9</td>
<td>28.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>11</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>V311(2)</td>
<td>HAM-D (n)</td>
<td>7.3</td>
<td>9.5</td>
<td>23.5</td>
<td>29.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>F313</td>
<td>HAM-D (n)</td>
<td>21.9</td>
<td>21.9</td>
<td>22</td>
<td>37.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>K317</td>
<td>HAM-D (n)</td>
<td>11.2</td>
<td>10.8</td>
<td>10.5</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>37</td>
<td>32</td>
<td>36</td>
<td></td>
</tr>
</tbody>
</table>

*IMI, NEW vs PBO, "p" less than 0.001
The dose-response study is a third kind of well-controlled study. If the effect of a drug increases with dose, that is evidence that it has an effect.

But it also gives information about what dose to use. E.g., you don’t want to use more than you need to get the full effect, because that could cause harm.
Good Dose Response

Until the late 1970’s, the uniform, standard, well-designed drug study titrated drug, to tolerance or effect, against placebo. The current standard, the randomized, parallel, fixed dose, dose-response study was never seen in drug development. Why?

They were used because:

• All patients got a high enough dose - “no waste” – an advantage in showing effectiveness
• Titration matches common practice
• It seemed a safe approach

But titration design give no useful information on D/R

• Designs confound dose effect and time effect
• The group on a particular dose is not randomly assigned to that dose; e.g., typically only poor responders receive the higher doses, leading to an apparent (but incorrect) umbrella-shaped D/R curve
• Because many diseases improve with time (or regress toward the mean), gives the impression that higher doses have greater effect
Epiphany and Examples

Chlorthalidone - Materson/Tweeddale

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.

<table>
<thead>
<tr>
<th>Dose</th>
<th>Fall in Blood Pressure (mmHg)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Supine</td>
<td>Standing</td>
<td></td>
</tr>
<tr>
<td>Placebo</td>
<td>0/2</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>12.5 mg</td>
<td>5/4</td>
<td>6/4</td>
<td></td>
</tr>
<tr>
<td>25 mg</td>
<td>11/5</td>
<td>15/7</td>
<td></td>
</tr>
<tr>
<td>50 mg</td>
<td>10/6</td>
<td>14/5</td>
<td></td>
</tr>
<tr>
<td>75 mg</td>
<td>11/6</td>
<td>14/6</td>
<td></td>
</tr>
</tbody>
</table>
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.
Impressed by Materson, who used a design we almost 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 E4
- In FDA regulations since 1985 as a kind of adequate and well-controlled study
- Frequent use, uniformly in hypertension, depression, and 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.
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)

<table>
<thead>
<tr>
<th>Adverse Experience</th>
<th>Assigned Treatment Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Placebo</td>
</tr>
<tr>
<td>N =</td>
<td>73</td>
</tr>
<tr>
<td>Dry Mouth</td>
<td>5</td>
</tr>
<tr>
<td>Somnolence</td>
<td>1</td>
</tr>
<tr>
<td>Asthenia</td>
<td>1</td>
</tr>
<tr>
<td>Dizziness</td>
<td>2</td>
</tr>
<tr>
<td>Headache</td>
<td>3</td>
</tr>
<tr>
<td>Impotence</td>
<td>1</td>
</tr>
</tbody>
</table>

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

No new law or disaster led to this third age, and it is now in full swing. Better D/R is part of this, but we have been moving swiftly toward closer attention to individual differences.

There has been growing recognition that individuals can have significant metabolic differences (some old cases, pseudocholinesterase, acetylation, but striking newer ones, like 2D6) and that other drugs being taken can have profound effects on metabolism, but evaluation of metabolism and interactions got a huge boost from a disastrous interaction (terfenadine), where inhibiting metabolism with CYP 400 3A4 inhibitors increased parent drug levels, causing QT prolongation and fata arrhythmias. It is now routine to examine these as well as age and gender differences in PK, to do population PK, and to examine effects of renal and hepatic impairment.
The Age of Individualization

PK differences are easily detected by blood level measurements. Pharmacodynamic differences between people were suspected (doctors pride themselves on individualizing Rx) but relatively few were documented and all of our clinical trials look at group, not individual, data, with little ability to look at individual D/R (unless x-over or forced titration, both unusual).

Genomics and proteomics change all this, and we’re just at the beginning of far more targeted therapy than we’ve ever imagined.
The Age of Individualization

Choosing a patient population in which it will be easier to show an effect is generically referred to as “enrichment.” With respect to showing effectiveness, there are two distinct possibilities:

- Individuals can differ in risk of an event (cancer, AMI); selecting high risk patients is “prognostic enrichment.”

- Individuals can differ in response to a treatment (e.g., because they have a particular receptor – low/high renin levels are a familiar example); selecting a responder population is “predictive enrichment.”

Such selection of people more likely to have an event, or more likely to respond to a treatment, makes it easier to demonstrate effectiveness. Let me give 2 examples.
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
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
Enrichment – High Risk Patients

1. Oncology (cont)


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.
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.
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.
Advantages of Predictive Enrichment

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

<table>
<thead>
<tr>
<th>Prevalence of Marker-Positive Patients</th>
<th>Response in Marker-negative Patients (% of marker positive response)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>100%</td>
<td>1.0</td>
</tr>
<tr>
<td>75%</td>
<td>1.8</td>
</tr>
<tr>
<td>50%</td>
<td>4</td>
</tr>
<tr>
<td>25%</td>
<td>16</td>
</tr>
</tbody>
</table>
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.