Better Drugs

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

Innovation in Health Care
Weill Cornell
October 27, 2012
Potential Improvements

A very broad subset, so today I will focus on a few areas of particular interest, particularly

1. Enrichment designs, including a few specific ones, such as studies in non-responders and randomized withdrawal studies.
2. Study simplification and benefit/risk considerations.

One overall note. Our practices over the last two decades in current regulations (21 CFR 312.80) and strongly encouraged by the new FDA Safety and Innovation Act (FDASIA), emphasize early discussions on optimizing study design and we have proposed, or will propose, guidance on adaptive designs, non-inferiority designs, dealing with multiplicity, and enrichment, all critical aspects of conducting studies that will do the job.
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 - Definition

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 than it would be in an unselected population.

Enrichment could also refer to a subset in a study that is to be used in the primary analysis, even if a broader population is studied.

The increased study power facilitates “proof of principle” (there is a clinical effect in some population) but it leaves open 1) the question of generalizability of the result and how the drug will work in other populations and 2) how much data are needed before or after approval in the “non-selected” group. (Do these patients benefit at all? Are they harmed?)

As will be noted, the main reason for enrichment is study efficiency – increasing the chance of success, often with a smaller sample size – but it also provides major benefits of individualization, directing treatment where it will do the most good and sparing people who cannot respond potential harm.
Enrichment

Enrichment is usually focused on effectiveness, but it is pertinent to safety.

- In the studies of oral hypoglycemics to rule out CV risk, we recognize the need to include high risk patients to have any chance at success in ruling out risk.

- One could show a drug lacks a class adverse effect by studying people who had the effect on another member of the class; enriching the population for likelihood of having the AE on the control and facilitating a showing of difference if there is one.
There are 3 broad categories of enrichment

1. Practical – virtually universal – decrease heterogeneity
   - Define entry criteria carefully
   - Find (prospectively) likely compliers (VA BP 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
Kinds of Enrichment (cont)

Apart from efforts to decrease variance, enrichment strategies fall into two distinct types:

2. 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 during the study. This is “prognostic enrichment.”

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 (e.g., ASA in primary vs secondary prevention).

3. Choosing people more likely to respond to treatment. This is “predictive enrichment.”

Choices could be based on patient characteristics, (pathophysiology, proteomic/genomic) or be empiric, based on patient history of response to similar drugs, early response of a surrogate endpoint (e.g., tumor response on some radiographic measure), or past response to the test drug (randomized withdrawal study).
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
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.
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.
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

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. As can be seen, the differences in event rates are huge, for both recurrence and survival.

Recent approval of MammaPrint, an in vitro test based on gene expression profile.
Prognostic Enrichment

2. Cardiovascular

Long routine to choose, in outcome studies, 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) study was in NYHA class IV patients. It needed only 253 patients to show a dramatic survival effect in a 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, needed only 4444 patients for a mortality effect. Later trials larger, used composite endpoints (i.e., not survival).
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.
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, 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

<table>
<thead>
<tr>
<th>Prevalence of Marker-Positive Patients</th>
<th>Response in Marker-negative Patients (0% of marker positive response)</th>
<th>Response in Marker-negative Patients (50% of marker positive response)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>75%</td>
<td>1.8</td>
<td>1.3</td>
</tr>
<tr>
<td>50%</td>
<td>4</td>
<td>1.8</td>
</tr>
<tr>
<td>25%</td>
<td>16</td>
<td>2.6</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, boceprivar and telaprilvir 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.
Predictive Enrichment – Empiric Approaches

There are many such possibilities. A few have been described:

1. Open trial followed by randomization
   - Oates, Woosley, Roden – anti-arrhythmic development
   - CAST: VPB suppression post-MI to prevent sudden death. Patients all screened for response; only randomized people with ≥ 70% VPB suppression
     Drug “worked” but was lethal
   - Beta-blocker CHF studies - screened for tolerability. Then withdrawn and randomized. Not a prediction of favorable outcome but of ability to tolerate

2. History of response to treatment class

3. Results in early studies of BiDil (isosorbide plus hydralazine) showed far greater response in blacks, allowing a definitive trial entirely in blacks

4. Adaptation: after interim look, include more of the responder population (e.g., men, disease severity); count everybody
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<sup>V600-E</sup>.

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, then are randomized to test drug (at one or more doses) or placebo. Endpoint can be time to failure (early escape) or conventional measure (attacks per week). Now standard for maintenance studies in depression and psychosis. Very recent NEJM showed recurrence of psychosis or agitation regression in patients with Alzheimer’s Disease who had responded to risperidone for 16 weeks and were then randomized to placebo.

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

- Open nifedipine
- Single-blind nifedipine
- Placebo

2 wk

4 wk

Randomization
<table>
<thead>
<tr>
<th></th>
<th>Nifedipine</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>Early withdrawal</td>
<td>0</td>
<td>5*</td>
</tr>
<tr>
<td>Early withdrawal plus AMI</td>
<td>0</td>
<td>6*</td>
</tr>
<tr>
<td>Investigator’s judgment of success</td>
<td>11</td>
<td>2*</td>
</tr>
<tr>
<td>Median angina/week</td>
<td>0</td>
<td>3.4*</td>
</tr>
<tr>
<td>Mean angina/week</td>
<td>0.7</td>
<td>18.4*</td>
</tr>
<tr>
<td>Change from baseline in attacks/week</td>
<td></td>
<td></td>
</tr>
<tr>
<td>better (\leq 1)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>same (\pm 1)</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>worse (\geq 1)</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>

*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).

We’re still discussing.
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

standard drug

standard drug
non-responder

new drug
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

<table>
<thead>
<tr>
<th></th>
<th>Clozapine</th>
<th>CP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGI (decrease ≥ 1)</td>
<td>71</td>
<td>37*</td>
</tr>
<tr>
<td>BPRS items (dec ≥ 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>concept disorganization</td>
<td>60</td>
<td>39*</td>
</tr>
<tr>
<td>suspiciousness</td>
<td>64</td>
<td>42*</td>
</tr>
<tr>
<td>hallucinations</td>
<td>59</td>
<td>51</td>
</tr>
<tr>
<td>thought content</td>
<td>65</td>
<td>40*</td>
</tr>
<tr>
<td>CGI and BPRS</td>
<td>15</td>
<td>2*</td>
</tr>
</tbody>
</table>

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


Lisinopril (ACE) 10 mg
Metolazone (diuretic) 1 mg
Losartan (angiotensin II antagonist) 50 mg

Patients with ACEI-induced cough
Taiwan and Hong Kong
n=84 elderly hypertensives, non-smokers

Lisinopril re-challenge 8 weeks, at least moderate
Placebo de-challenge 4 weeks, not at all

Randomize to 3 drugs, 10 weeks
Assessment by questionnaire, nurse
## Cough

<table>
<thead>
<tr>
<th>Drug</th>
<th>Cough rate, any</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lisinopril</td>
<td>97%</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Losartan</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>Metolazone</td>
<td>21%</td>
<td></td>
</tr>
</tbody>
</table>

*Note: The cough rate data is statistically significant, indicated by p < 0.001.*
Other Potential Improvements

By far the most important steps are those related to study design and choice of population, all considered above under the heading of enrichment.

But there are other steps to mention

1. Adaptive designs

2. Efficiency
Cost

If a drug is eventually shown effective and is marketed, people probably don’t worry too much about cost. But there are worries

- Cost can affect the decision to try at all, e.g., with a novel drug (i.e., no track record, no history) in a setting that needs a large study.
- Cost can alter the studies being done, e.g.
  - Discourage a promising line of testing
  - Lead to poor decisions, such as inadequate dose finding

So efficiency does matter, and it has 2 main aims:

1. More efficient studies
2. A greater likelihood of success
More Efficient Studies

1. Making the data easier to collect or collecting less of it.

There has been a great deal of discussion of over-extensive data collection, especially in late studies, and draft FDA guidance has suggested a “phase ¾ lite” approach, not collecting non-serious AE’s that don’t affect therapy after there is a sizable database already, not collecting all concomitant therapy, even reducing the scale of baseline history. Some of this is not new.

- The large simple trial is based on the idea that the whole population is of greatest interest, so too much data on baseline variables, and too much selection, is a distraction. Greatly reduced CRF size was an important goal.
- Cardio-renal has long urged less collection of minor AEs and concomitant therapy
- Certainly a great hope is that baseline data, and perhaps outcome data, can be obtained substantially from existing patient records
- EVERYONE agrees that current informed consent practices are ineffective, hugely time consuming. We have seen attempts to do this online.
Dose-Response

We have made great progress since the 1980’s in doing randomized, fixed dose, dose-response studies (ICH E-4 endorses them). This came about at least partly because doses of diuretics in the 1960’s and 1970’s were 100 mg of HCTZ and chlorthalidone, causing significant hypokalemia and, probably, CV death, but giving no increased effect compared to lower doses. We also had no idea what the right dose of digoxin or reserpine was, resulting in fatal digoxin overdose, and excess depression with reserpine.

But we’re not all better.
More recently, a potential blockbuster drug for IBS associated with diarrhea, Lotronex (alosetron) was removed from the market because of cases of ischemic colitis and severe constipation leading to fatal obstructions. It has been returned under a limited distribution with a reduced dose. The entire phase 3 program used a 2 mg daily dose, producing severe constipation in about 1/3 of patients (this for a treatment for diarrhea). In retrospect, it seems clear that doses of 0.5, 1, 2 mg, or dropping back after an initial 2 mg or going to q 2d treatment should have been tried. The drug was returned to market with a recommended dose of 1 mg.

A similar error, leading to withdrawal, was made with astemizole, a long half-life anti-histamine; it should have had a loading dose with subsequent 2-3 mg doses, instead of the recommended 10 mg daily dose. Then it would probably not have caused TdP.
General advice on D/R

ICH E-4

1. Strong encouragement to identify whole D/R curve for benefit and toxicity; frequent error is too narrow a range. Placebo desirable. Dose-finding is NOT over in phase 2.

   • Choose starting dose
   • Identify titration steps
   • Find plateau-dose increase useless
   • Don’t forget dose interval

2. Group values are what we get but individual D/R also of interest. Need to give each person > 1 dose to find this, rarely done, but it is possible using forced titration or optional titration with NONMEM or other analysis. These designs, while more complicated than randomized, fixed dose D/R, probably deserve more attention.