Impact of Comparative Effectiveness on FDA

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

Comparative Effectiveness Research Symposium
UCSF
Jan 29, 2010
Current Status of CE at FDA

I. Legislative history

The 1962 amendments require for approval substantial evidence that a drug will have its claimed effect, with the source of that evidence being adequate and well-controlled studies.

The Senate report made it very clear that there was no relative effectiveness requirement – a new drug need not be better than, or even as good as, available therapy.

With an exception I’ll note, we have followed this practice and, indeed, we usually do not have direct comparative data in placebo-controlled studies.
II. An Important Exception

We have clearly noted that if inferiority is a matter of life and death (or major morbidity) we can and do consider it.

This was stated clearly in April 1995 in a “Reinventing Regulation of Drug and Medical Devices” document by President Clinton and VP Gore, which was also set forth in the August 1, 1995 FR as an FDA “position,” signed by Bill Schultz, Deputy Commissioner for Policy.

Exact status of this statement is not so clear, but you’d think it would be a “guidance equivalent.” It says some interesting things.
The 1995 notice was intended to “address concerns about a comparative effectiveness standard” that had been raised by drug and device manufacturers.

It said that in general, “for most new drugs and devices intended to treat serious illness or provide symptomatic relief, effectiveness is shown by comparing the drug to placebo; i.e., there is no comparison to another active treatment.

“In certain circumstances, however, it may be important to consider whether a new product is less effective than alternative therapies, when less effectiveness could present a danger to the patient or to the public,” e.g.
CE and FDA, Exception (cont)

1. When the disease is life-threatening or causes irreversible morbidity, or
2. Contagious disease

The notice also said we are willing to approve drugs for subpopulations who do not respond to, or don’t tolerate, available therapy, a kind of comparative effectiveness.

We do, of course, and can, accept greater toxicity in that case, but note that effectiveness in non-responders is rarely shown properly, i.e., by randomizing non-responders back to the failed or intolerant treatment, and to the test drug. This is a matter for another day, but I’m aware of only 3 successful studies in non-responders, clozapine, bepridil and captopril, and one total failure, VIOXX in Celebrex non-responders.
Note that without a celecoxib control, rofecoxib would have appeared VERY effective in this NR population.
So, in practice, we clearly expect comparable effectiveness of a new drug for life-threatening or debilitating illness, such as cancer or serious infectious diseases.

In practical terms, we ask for comparable effectiveness whenever it would be unethical to do a placebo-controlled trial because it would mean denying patients a known effective therapy that would prevent death or serious disability (clearly stated in ICH E-10).

In those cases non-inferiority (NI) studies, a kind of active control trial, are carried out to show effectiveness.
NI studies show effectiveness by showing, at a minimum, that a new drug is not worse than the active control by an amount (the NI margin) equal to the whole effect of the active control; i.e. some of the effect is preserved. The effect of the active control is called $M_1$, the largest possible NI margin. A difference between the new drug and control ($> M_1 = $ all effect lost).

BUT, if you must ethically do an NI study, the control effect is of perceived value, so we usually ask that some fraction of the control effect be preserved (typically 50%), so the study needs to rule out loss of $M_2$ (the fraction of the whole effect of the active control that must be preserved).

This IS a kind of relative effectiveness requirement.
III. Comparative Effectiveness Claims

There are 2 kinds of claims to consider:

- Similarity/equivalence
- Superiority

When an NI study succeeds in showing non-inferiority (i.e., that at least 50% of the control effect is preserved), that does not show equivalence, or support a claim of equivalence, which would be a much higher standard, like the 80-125% CI we demand for bioequivalence for generics.

I can’t think of any cases where we’ve accepted such a claim other than for some topical products (skin, eye) that did indeed meet the 80-125% standard.
Superiority claims have been sought and our standard has been the approval standard: adequate and well-controlled studies (usually more than 1). Moreover, the studies must be fair, as discussed in ICH E-10 [Choice of Control Group and Related Issues in Clinical Trials, 2001]. A comparison could be unfair if:

- Low dose of the comparator was used.
- The patient population had previously failed the older drug.
- Selection and timing of endpoints favored one drug.
Comparative Claims (cont)

It is not easy to get such a claim, but there have been successes in oncology and elsewhere.

- Two large studies showed candesartan had a larger effect than losartan (in labeling).
- LIFE study (losartan vs atenolol) showed superiority vs stroke, but in only one trial. Losartan got stroke claim.
- Clozapine effective in haloperidol and thorazine non-responders
- PPIs have claims vs H2 blockers
- Anastrazole is superior to tamoxifen as adjuvant Rx post surgical treatment of breast Ca, especially in ER positive.
- Irbesartan delayed decline in renal function in type 2 diabetes; superior to amlodipine, which had no effect.
Comparative Claims (cont)

We thus use the legal effectiveness standard for what is, in fact, a claimed effect, just as the law demands. It is a high standard, but it is not easy to see how a lesser standard would fit the law and (my opinion here) whose interest such a standard would serve.

We know that before there was an effectiveness standard, the effectiveness of thousands of drugs and more thousands of claims were unsupported and proved unsupportable. We know that claims for diet, any supplements, unencumbered by any requirement for controlled studies, are universally unsupported. It is not easy for me to see a public interest in the proliferation of comparative effectiveness claims based on data known to be unsuited to the purpose.

I will elaborate (note, by the way, that safety differences may be so large that methods other than clinical trials might detect them).
IV. Current Interest in CE

As anyone can see, there is a large and growing interest in CE, and, of course, there should be.
Comparative Effectiveness

The excitement is palpable. . . And why not?

Clinically, after knowing a drug works and is safe (which FDA takes care of) most of the important questions about drugs are comparative, i.e., deciding which drug to choose:

- Does it work better than alternatives? Faster?
  - In all patients
  - In a subset
- Can you add it to other treatments?
- Does it have some additional benefit in some or all patients?
- Does it work when others fail?
- Is it about as good, but cheaper?

But there usually isn’t much of such data:

- Drug companies historically have not done proper comparisons (with the critical exception for situations where active control trials are ethically necessary)
- Trials almost never have > 1 comparator; usually interest is in comparing all members of a class
- Trials rarely compare across classes
- Trials usually are too small to give definitive answers
So the medical need for comparative data is great and apparent.

We also need to acknowledge a major interest in costs of therapy. All of us, payers too, will pay for a more expensive treatment with an advantage
- maybe after other therapy fails
- maybe it depends on how much advantage
but there is great reluctance to pay more for the same effect. So a major interest of payers is showing whether there is an advantage. (Could they just agree to pay only when one is shown?)

But wanting comparative data does not necessarily mean we know how to get comparative data of high quality with reasonable effort and at acceptable cost.

And it must be of high quality. Mistakes will greatly undermine the credibility of the effort, not to mention the harm they could do.
Comparative Effectiveness Is Not the Only Need

I realize there is current enthusiasm for comparative effectiveness, but we need to keep our balance. If there is to be funding for trials there are other critical issues too. For example:

1. Do our physical therapy and non-pharmacologic psychiatric interventions work at all? Many are untested.
2. How can we improve compliance/persistence with vital chronic therapy (lipid-lowering, BP, diabetes control, smoking cessation, weight reduction)? Could cluster-randomized trials help?
3. How low should we push LDL, BP, BS; is it the same for everyone? How many anti-platelet treatments should we give in ACS and after PCI and how long should we give them?

The right determination of what to study is the value of what we’d learn, not whether it is comparative. The best study may be a comparison of a treatment with no added treatment. The IOM list of 100 is very consistent with this.
Comparative Effectiveness Issues

Comparative effectiveness raises a host of issues, all of them interesting and most of them matters of long FDA and personal interest, including:

1. How we obtain evidence of comparative effectiveness and safety: role of trials, meta-analyses, observational data.

2. Often (usually) you’re interested in comparisons with multiple drugs, not just one, frequently drugs in different pharmacologic classes. How to compare multiple treatments is challenging and doing it is costly.
3. There are major challenges in doing comparative effectiveness trials

- Differences between effective treatments will, at most, be small, so that
  - Trials will need to be very large to show them
  - Nothing but an RCT is credible

- Showing there is no (or not much) difference between treatments, often the goal of the comparison, is also very hard, will often need a placebo group to assure assay sensitivity, and again, trials may need to be very large, depending on the size difference to be ruled out: the non-inferiority study problem

- Efficiency and simplification are critical
Comparative Effectiveness
You Need Randomized Trials
(Maybe Meta-analyses)

With rare exceptions, differences between drugs, if any, will be small, considerably smaller than the whole effect of the drugs, which themselves are often small. And the difference you want to rule out is also small.

A blockbuster outcome study in CHF, hypertension, CAD will reduce event rates by 40%. Far more commonly, it will be more like 20%. That is the whole effect of the drug, an HR of 0.8. A complete loss of that effect (1 / 0.8) would give an HR of just 1.25 for the comparison of a new drug vs the standard; i.e., it would be only 25% worse.

But between-treatment differences of interest, or the difference to be ruled out, will be smaller than that; suppose you would want to detect a loss of half that effect, a 10% difference. In that case the HR for the inferior drug, the upper bound of the CI for new/old, would be just 1.125, i.e., very hard difference to detect.

In terms of risk, that means you’re trying to detect a risk ratio of 1.1-1.2 at most. This is possible in large ambitious RCTs, but you cannot reliably detect such differences in anything but randomized trials.
Symptomatic conditions pose at least as great a problem, at least usually (and one might ask how important it is to rule out or document small differences).

Trials of antidepressants fail about 50% of the time (cannot distinguish drug from placebo) and a typical effect size is 3 HamD points (drug-placebo). Trials these days are 100-200/arm.

A large between drug difference could conceivably be 1.5 HamD points (that would be a very large difference and, usually, the less effective agent would have had difficulty beating placebo). Far more likely would be a difference of 1.0 HamD points.

Trials to show such differences would be enormous. Moreover, failing to show a difference would be meaningless without a placebo group to assure assay sensitivity (ability of the study to detect effects).

Most symptomatic conditions are like this, except where effects are huge (Tysabri vs interferon, a difference so large it is obvious in cross-study comparisons).
Comparative Effectiveness
You Need Randomized Trials (cont)

It is not insulting to observational/epidemiologic approaches to say that they are generally unreliable when trying to detect risk ratios of < 1.5, and certainly when looking for risk ratios of 1.2 and less. It is not a lack of power. What makes such approaches tempting is in fact their huge power and speed.

But those advantages do not make up for potential bias and confounding. There are many sobering examples. Let me give two:

- Hormone replacement therapy
- Calcium channel blocker toxicity

The incorrect results, of epidemiologic studies in these cases, unfortunate at best, disastrous at worst, did not usually arise from obvious methodological flaws or foolishness. The methods are just not reliable for small differences, usually because without randomization you cannot assure the needed close similarity of the groups receiving each treatment.
Hormone Replacement Therapy

Although observational studies did not give uniform results, hormone replacement therapy was thought to reduce coronary heart disease (CHD) by 40-50%. The Women’s Health Initiative randomized > 16,500 post-menopausal women 50-79 to HRT (0.625 oral equine conjugated estrogens + 2.5 mg medroxyprogesterone acetate) or placebo.

Despite favorable effects on LDL and HDL cholesterol and triglycerides, coronary heart disease effects were adverse.

<table>
<thead>
<tr>
<th></th>
<th>HRT 8506</th>
<th>Placebo 8102</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD</td>
<td>188</td>
<td>147</td>
<td>1.24</td>
<td>1.00-1.54</td>
</tr>
<tr>
<td>NFMI</td>
<td>151</td>
<td>114</td>
<td>1.28</td>
<td>1.00-1.63</td>
</tr>
<tr>
<td>Fatal CHD</td>
<td>39</td>
<td>34</td>
<td>1.10</td>
<td>0.70-1.75</td>
</tr>
<tr>
<td>CHD, revasc, angina</td>
<td>369</td>
<td>356</td>
<td>1.00</td>
<td>0.86-1.15</td>
</tr>
</tbody>
</table>

The risk was increased within the first year.
HRT

HRT has obvious short-term benefits but the case for CHD prophylaxis, logical (women have less CHD than men while producing hormones and catch up with men after menopause) and epi-supported, was not only not made, but CHD harm was strongly suggested.

There were also increases in breast Ca, thrombophlebitis, pulmonary emboli.
Calcium Channel Blockers

The full CCB story deserves a book, not a few slides. Over the course of several years, roughly 1995 through 2002, cohort and case control studies, almost all of them comparing CCB’s with other antihypertensive drugs, suggested that CCB’s:

1. Increased the rate of AMI (Psaty, et al, JAMA, 1995).

2. Increased mortality (Furburg and Psaty, Circulation, 1995) actually a subset of a meta-analysis of nifedipine).

3. Increased mortality (Pahor, et al. J Am Geriatrics Society, 1995, a cohort study). Oddly, verapamil was protective; diltiazem, nifedipine AND ACEIs all gave RR’s of 1.5-1.9.
Calcium Channel Blocker (cont)


5. Increased risk of all cancer (Pahor, et al Lancet, 1996). Oddly, risk was up for verapamil and nifedipine, not at all for diltiazem.


Calcium Channel Blockers

FDA’s Cardio-Renal AC saw the mortality and AMI data (probably in 1995-6) and did not find it persuasive. HRs were mostly in the 1.5-2 range and varied considerably from drug to drug.

To my best knowledge, none of these findings were confirmed in RCTs (ALLHAT, various CAD trials of verapamil and diltiazem). The findings were discussed, condemned, supported in dozens of papers. A Sounding Board piece (NEJM) in 1997 by Deyo, Psaty, and others described manufacturers’ attempts to gain access to Psaty’s records related to the 1995 AMI study, as well as many hostile academic (perhaps manufacturer-supported) critiques, citing this as a classic case of attacking scientific results that run counter to financial interests and strongly-held beliefs. That surely could be part of it but there were certainly scientifically sound bases for criticism as well. Paper (can’t find) comparing industry support for authors supportive and opposed to the CCB findings. Guess which ones had more support.
Calcium Channel Blockers

People can form their own views as to what all this illustrates. Among other things it shows

1. Inadequate attention to description and presentation of epi results. Epi studies need careful protocols that record changes, well-described hypotheses, correction for multiple hypotheses (i.e., all the things we’ve learned to ask about RCTs).

2. Particular risks when the adverse effect is a possible consequence of the disease, where the severity of the condition and the effect of treatments can be confounded.

3. RR’s < 2 need great care and should be viewed very skeptically (although they can surely generate hypotheses). Comparative effectiveness will almost invariably be about RR’s < 1.5 (whether or not they find a difference).

4. Epi errors can cause major disruptions.
Calcium Channel Blockers

With recognition of the need to get BP under better control, CCB’s must be used in many people. They may even have advantages in some populations. But their use was somewhat marginalized for many years because of these concerns. There is little of that concern expressed in JNC VII (2004), so perhaps the damage has passed.

WE DO NOT WANT ERRORS. The questions addressed in comparative studies, especially outcome studies, matter. To get correct answers, the comparisons need RCTs unless differences are very large. They hardly ever are.
Comparative Controlled Trials - Difficulties

There is not a great deal of experience in doing such trials properly, and the challenges are substantial.
Comparative Effectiveness - Difficulties

I. Multiple Drugs of Interest

What physicians really want to know is how all (or at least many) members of a class compare. This is not easy, for many reasons.

1. For many comparisons you need a placebo to assure assay sensitivity, a potential problem for post-approval, often large, studies.

You can sometimes use a NI study design where there is a solid basis for knowing the effect of the positive control in an NI study, but that would be impossible in depression, anxiety, and most symptomatic conditions; for those you need a placebo to show ASSAY SENSITIVITY, i.e., that you can tell one thing from another, because many studies in those conditions cannot tell active drugs from placebo [You could show superiority without the placebo, but not similarity].
Comparative Effectiveness – Difficulties

2. Hard to expect a company to study multiple drugs in one study.

Separate comparisons don’t really tell you what is needed; you can’t usually compare across studies.

Multiple comparisons by government

- **ALLHAT** – chlorthalidone, lisinopril, doxazosin, and amlodipine

Ambitious but results hotly debated; there were design problems (couldn’t add diuretic to lisinopril). Meta-analyses suggest different answers.

*Did* show that cheapest drug (chlorthalidone) was a reasonable start, but drugs have different properties: some treat diabetic nephropathy (ARBs), CHF (ACEI’s, BBs, diuretics), angina (CCBs, BBs), or post-infarction (BBs, maybe ACEI’s).
ALLHAT

Wonderful Intent, Hard Trial

Compared – clorthalidone, lisonopril, amlodipine, and doxazosin for... effectiveness.

Some element of cost: “Are newer types of anti-HT, which are currently more costly... as good as or better than diuretics in reducing CHD incidence and progression” (abstract, Am J HT, 1996; 9: 342-360).

Problems:

1. Plainly, an NI study, but no discussion if NI margin for any endpoint. Doing so would have been difficult because regimens did not match past effective regimens and population (enriched for black patients) not the same (35% black). Did this disadvantage lisinopril? The question is, then, what does failure to see a difference mean? It is very hard to know and, to my best knowledge, not addressed at all.
Problems (cont)

2. No beta-blocker group.

3. Treatments did not get usual accompaniments because you could not add another test drug.

   E.g., could not add diuretic to lisinopril. This is particularly critical for black population and for CHF (all CHF studies of ACEI’s were added to diuretic). Lisinopril thus had slightly poorer control of BP.

4. ACE inhibitors were superior for CV events in a different study, the Second Australian National Blood Pressure Study (HCTZ, mostly white).

5. Did we learn enough? I’d say yes: main lesson is that it doesn’t matter too much how you get the BP down.
Comparative Effectiveness – Difficulties

- **CATIE**

  NIMH: 4 atypical (olanzapine, risperidone, quetiapine, ziprasidone), one typical (perphenazine) anti-psychotics used in schizophrenia showed olanzapine was most effective (fewest D/C for lack of effectiveness) and least well-tolerated (most D/C for intolerance). CATIE worked because there were differences. Had there been no differences, it would have, absent placebo, been wholly uninformative.

  Both ALLHAT and CATIE were very expensive. Perhaps worth it but at those prices can’t do too many.
1493 schizophrenics randomized to olanzapine, perphenazine, quetiapine, or risperidone (later ziprasidone).

Endpoint was “discontinuation” of treatment for any cause.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Olanz 330</th>
<th>Quet 329</th>
<th>Risp 333</th>
<th>Perph 257</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All DC (%)</td>
<td>64</td>
<td>82</td>
<td>74</td>
<td>75</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Lack of E (%)</td>
<td>15</td>
<td>28</td>
<td>27</td>
<td>25</td>
<td>0 signif &lt;</td>
</tr>
<tr>
<td>Intolerability (%)</td>
<td>19</td>
<td>15</td>
<td>10</td>
<td>16</td>
<td>0 signif &gt;</td>
</tr>
</tbody>
</table>
II. Sample size is very large

Suppose you wanted to compare anti-depressants. Current studies vs placebo these days use 100-150 patients per group to show a drug-placebo difference of 3-4 HamD points. You need placebo for assay sensitivity. What HamD difference do you want to rule out?

2 points – no chance it’s that large
1 point – sample size for active drug would be many hundreds, perhaps 1000. Is that really feasible?
Comparative Effectiveness - Difficulties

Given the problems (multiple drugs of interest, small effect sizes) it is tempting to seek alternative data sources, notably meta-analyses and cross-study comparisons. These are treacherous. In a cross-study comparison patients are not randomized to treatments and patients on one drug may differ from patients on another, making such comparisons notoriously unreliable. The problems and potential biases in meta-analyses are well-recognized and only rarely will there be a sufficient number of trials to pool.
Possibilities

The problems I’ve described can perhaps be overcome, if there is enough interest. Possibilities include

- Doing large studies in treatment environments already collecting data (HMO’s, VA), perhaps using internet to enroll, gain consent, follow PRO outcomes. Do not select too much. (I’m describing very pragmatic trials). We know very large trials in Europe (ISIS, GISSI) had reasonable costs.

If patients and doctors were “into” this, maybe it wouldn’t cost too much.

- Placebos are hard in the real world but you don’t need one to show superiority. But in symptomatic conditions, absence of a placebo will lead to inability to interpret results if no treatment is superior.