

# Interpretation of Observational Studies of Cardiovascular Risk of Nonsteroidal Drugs: Richard Platt MD

## JOINT MEETING OF THE ARTHRITIS ADVISORY COMMITTEE AND THE DRUG SAFETY AND RISK MANAGEMENT ADVISORY COMMITTEE

February 16-18, 2005, Hilton Gaithersburg, 620 Perry Parkway, Gaithersburg, Maryland.

Highlights.....	1
Presentation Text.....	4
Presentation Slides .....	14

### Highlights

#### OBSERVATIONAL STUDY LIMITATIONS:

- **APPLES AND ORANGES:** In observational studies one can never be sure that the groups being compared are comparable in risk for the outcome of interest. One can only adjust for the confounders that one knows about, and the presence of other confounders can never be excluded. If the difference in estimate of risk between the unadjusted and the adjusted result using known confounders is small, one can have somewhat more confidence in the result. Similarly, the presence of a dose-response relationship imparts more credibility.
- **IS SYSTEMATIC BIAS A REASONABLE EXPLANATION FOR THE RESULTS?** Since differences between drug groups may reflect factors unrelated to the drugs being compared, one may be tempted to conclude that the magnitude of the apparent effect is so large that any systematic bias could not be enough to explain the findings. However, the estrogen “protection” story in post-menopausal women provides a cautionary tale.
- **DIFFERENTIAL PROBABILITY OF ASSIGNMENT TO DIFFERENT DRUGS:** Political, economic, standard-of-care or geographic factors may influence the probability of assignment to different drugs.
- **DIFFICULT TO MONITOR DRUG ADHERENCE AND DATA COLLECTION:** identifying adherence to treatment and making sure that events and other data items are properly collected is often difficult.
- **IMPACT OF PRIOR DRUG THERAPY:** It is difficult to exclude an impact of prior drug therapy on outcomes occurring on subsequent therapy.

- **FOLLOW-UP OFTEN INCOMPLETE:** In cohort studies, initial drug status may change during follow-up and adequate follow-up is difficult.
- **EXTRAPOLATION TO DIFFERENT POPULATIONS:** One has to be careful about extrapolating from the population you study to the other populations or the general population with the disease. Case-control studies may not provide patients who are representative of the “group that you are trying to study” although this is less of a problem in cohort studies. The outcome to be evaluated may be limited by the need to get information on drug therapy from a patient surviving the outcome of interest.
- **BIAS:** Recall bias and reverse recall bias is a common problem.
- **LESS GOOD FOR COMMON EVENTS:** Common events are harder to study in observational studies than rarer events.
- **LACK OF STANDARDS FOR OBSERVATIONAL STUDIES:** The lack of uniform standards for designing, documenting, analyzing and interpreting observational studies is a major problem.

### **OBSERVATIONAL STUDY ADVANTAGES:**

- **QUICKER & CHEAPER:** Observational studies are quicker, cheaper and less resource-intensive.
- **MORE REPRESENTATIVE OF POPULATION ON DRUG:** Subjects in some types of observational studies are more likely to be comparable to the actual population of users for the drug and to use the drugs as they are used in the general population.
- **GOOD FOR RARE EVENTS:** They are particularly useful when evaluating rare events.

### **CLINICAL TRIAL LIMITATIONS:**

- **SLOW & EXPENSIVE:** Clinical trials give slower answers than observational studies and can be very expensive and resource-intensive.
- **NOT REPRESENTATIVE OF POPULATION ON DRUG:** Subjects in clinical trials are “different from the actual population of users” of a drug.
- **LESS GOOD FOR RARE EVENTS:** For very rare events clinical trials may not be practical.
- **LACK OF STANDARDS FOR CLINICAL TRIALS:** Although clinical trials standards are better documented than the standards for observational studies, deficiencies remain in 1) adjusting p values for multiple testing and other factors, 2) exact pre-specification of primary and secondary hypotheses, 3) documenting and displaying the results of analysis, and 4) designing

and documenting valid trial stopping

rules.

## CLINICAL TRIAL ADVANTAGES:

- **STUDY GROUPS MORE LIKELY TO BE COMPARABLE:** The “tremendous advantage” of randomized trials is that the different groups being compared have a known probability of being within certain limits of comparability at baseline, without systematic bias in group assignment. However, even in properly randomized trials, major baseline inequalities can sometimes happen by chance.
- **RANDOMIZED TRIAL MORE CREDIBLE:** Overall, “all things being equal ....a randomized trial is

more credible ..... than an observational study.”

- **GOOD MONITORING OF DRUG ADHERENCE AND DATA COLLECTION:** Randomized trials are better at identifying adherence to treatment and making sure that events and other data items are properly collected.

## OTHER ITEMS:

- **OBSERVATIONAL STUDY DESIGNS:** If observational studies are to be done to evaluate a common event, case-control, nested case-control or cohort designs are best. For the present COX-2 issue, Kimmel was case-control, Graham and Solomon were nested case-control, and Ray and Aramis were cohort designs.
- **COHORT STUDIES:** A cohort study identifies if a patient has been exposed to the drug or not and then follows the patient to record an outcome. Inception cohorts are people who had to be members of a population (e.g., members of a particular health plan) from its inception (or for a defined period before beginning the drug of interest).

- **CASE-CONTROL STUDIES:** A case-control study starts with people who have the outcome we care about (e.g., MI) and matches them to comparable patients who did not have that outcome.
- **NESTED CASE-CONTROL STUDIES:** A nested case-control study is a “hybrid” that “draws many of the strengths from both designs” (cohort and case-control). It is a case-control study that is “nested” in a defined population, thus combining the strength of a cohort study with the efficiency of a case-control study.
- **REASONS FOR LACK OF POWER:** As with randomized trials, observational studies may have insufficient events, insufficient duration of therapy, or not-at-risk

patients that prevent identification of the outcome of interest.

- **INTERPRETATION OF RISKS:** “I think that observational studies are best at finding relative risks that are more than 2. I think that I would pay some attention to relative risks of 1.5. I get very nervous about adjusted relative risks of 1.2.” Calculating other risk estimates such as absolute risk, person-level risk and population-level risk is also important when making individual patient-physician decisions, and in making public policy decisions.

- **APOLOGY:** Note that some of the “Highlights” above were derived from related comments made by other committee members during the COX-2 meeting. It was felt to be important to try to draw together some of the common threads comparing observational studies and clinical trials. The reader’s (and Dr. Platt’s) forbearance is requested. This is the only “Highlights” section in the COX-2 meeting summary in which this approach has been taken.

## Presentation Text

Thanks. The framers of the meeting thought it would be useful at this point to have a discussion about observational studies to put us all on the same page.

There was a view by some that the expertise around the table might be uneven and it would be worthwhile to have some discussion about some of the basics. It is clear that that is not the case.

I realize that a number of the people here have written a book and several of my teachers are here, so to that extent, I think we can either make this a quick discuss or use this as an opportunity for a real interactive discussion, because there are some hard questions here and no matter how we sort we out, we are going to be left with less than in the way of firm answers than we would like.

I also understand that there is a point of view that says that there are lies, damn lies, and observational studies, so part of what I think is worth doing is using this

time maybe to take our temperature about whether and under what circumstances we can put weight on observational studies.

We saw a version of this slide last night actually in the last presentation about why perform observational studies at all, because I subscribe to the general view that all things being equal, a clinical trial, a randomized trial is more credible, provides more information than an observational study.

The problem is all things aren't always equal and so there are reasons to ask what we can learn from observational studies.

I think the most important of them is no matter how well a clinical trial is designed, the individuals who are recruited and consented to a clinical trial are inherently going to be different from the actual population of users, and if we want to understand how an agent performs among real users in the way

they actually use the drug, then, I think there is no escape but to look to observational studies.

Additionally, observational data is by definition there, so when a pressing question arises, sometimes observational data is the first way we can get insight into the relationship between the drugs we care about and the exposures.

I think in that regard, these studies can often be thought of as helping us identify the areas in which it would be most fruitful to invest in full-blown randomized trials. We will never live in a world where we are able to do all the randomized trials we care about.

I know that Charlie Hennekens' landmark randomized trial of aspirin was preceded by, as I recollect Charlie, a large number of observational trials, it made you think that it was reasonable to do those randomized trials, so observational studies can be useful in that regard.

Finally, when we are talking about trying to understand effects that are relatively unusual, we stress even the largest clinical trials. We talked yesterday about the fact that the most recent drug approvals have used much larger populations in the NDA phase than had been studied in the old days, and yet they are still small compared to the numbers needed to parse out relatively small differences.

There are a lot of different kinds of observational trials. I have listed a few of the most common. The ones between the lines here are the ones that are really the subject for discussion here.

Tom Fleming made the absolutely correct and somewhat counterintuitive point that it is often more difficult to do good observational studies of relatively common outcomes than rare ones, and because of that, the group of studies that I think at least are reasonable to consider for looking at relatively common outcomes are case-control studies, nested case-control studies and cohort studies.

We have examples of each in the materials that have been handed to us. The study by Kimmel is a pretty traditional case-control study. The studies by Ray are cohort studies, as is the Aramis study. The study by Dave Graham and the Solomon study are nested case-control studies.

Just as a quick reminder, the distinguishing feature of cohort studies is the fact that the study population is defined on the basis of whether people are exposed to the drug or not, and then we look forward to what happens to them. In that way, they are exactly comparable to clinical trials, with the big difference that the assignment to drug is not randomized.

The strengths of those compared to case-control studies are you have a reasonable shot at the outset of selecting individuals who are representative of the group that you are trying to study, and if you organize the study properly, you have a reasonably good chance of getting unbiased exposure assessments.

The weaknesses, particularly of observational cohort studies is that just because individuals had the right drug exposure at the outset, they may change that. You can deal with that with an

intention-to-treat design, but you pay for a price for that, and in observational studies, loss to follow-up is a big problem.

We are particularly plagued by that because the large majority of the observational studies we are working in are ones that use administrative data from one sort of health plan or another, and individuals move in and out of health plans, so that it becomes difficult to follow them over time.

Case-control studies, remember are ones that start with individuals who have the outcome we care about, myocardial infarction or myocardial infarction and sudden death, and compares them to individuals who haven't had that experience, then, you look back and ask what their drug exposures are, the reasons for doing those studies are that they are, first of all, very efficient studies.

You don't have to study thousands and thousands. You can study as many cases as you find and a reasonable number of controls, and you can look back and classify exposure however is most useful, and that is a very convenient and versatile feature of case-control studies.

The big weaknesses are that it is very hard to assure oneself that the cases and the controls are really representative of the populations that you care about, and for conventional case-control studies, for instance, the study by Kimmell that we are going to look at, it takes a lot of work to be sure that people who know what they have already experienced an MI don't differentially report their exposure to the drugs that we care about.

That can be for all sorts of reasons and it might not even be wrong, but the individual who has had an MI and might be just thinking harder about whether he or she had been exposed to a drug that we care about.

By the way, nested case-control studies, for instance, the study that David Graham did is a hybrid that really, in my view, draws many of the strengths from both designs, that is, because nested means the case-control study is nested in a defined population, so it has a lot of the strengths of cohort studies and some of the efficiencies of the case-control studies.

The differences between the observational studies and randomized studies are pretty clear. Randomized trials have the tremendous advantage that there is lots more reason to expect the treated and untreated groups to be comparable to one another.

There is a lot more opportunity to be sure that the outcome assessment and adherence to treatment are good or at least well known, and we have reviewed the difference for the observational studies.

I think it is worth making the point that there are a substantial number of similarities between observational and randomized studies. Just because we randomize individuals in randomized studies, it doesn't mean that the treated and untreated groups are comparable.

We talked about a study yesterday that was a randomized trial where there was a substantial imbalance in important risk factors. So, it is incumbent no matter what kind of study you do, I think to

look for comparability, and both studies have as potential weaknesses that there are risks of false positive results and doing subgroup analyses and multiple comparisons increases that risk.

We talked a fair amount about that yesterday, and both are at risk for false negative results. That can be partly because the studies may not be powered well enough either because there is insufficient sample size or individuals aren't studied for a long enough duration to see the biological effects that we care about, or a vulnerable group just isn't included.

That is a problem with both kinds of studies and I think all studies have to be evaluated on their own merits, so let's just step through the various places where observational studies might be into trouble or at least the things that need careful assessment when we look at these studies.

The first is: Are we studying the right outcomes? It is essentially impossible in any of these observational studies to use the kind of rigorous adjudication that is a hallmark of the randomized study, so I think we are going to have to ask ourselves are these outcomes good enough.

The several kinds of outcomes in the studies that we have been asked to look at are hospitalized MIs. The case-control study by Kimmel uses survivors. It had to use survivors because they were collecting the exposure information by interview after the individuals had left the hospital, so if we care about all MIs, then, that study isn't going to tell us what we want to know.

Some of the studies use MI and out-of-hospital sudden death by linking to vital statistics records. I think that is probably the closest we can get in observational studies to the intention-to-treat all outcome designs of the randomized trials, and some of the studies use composite designs.

You have to ask are these outcomes measured appropriately. Most of the studies that we are looking at use some form of automated medical record or claims data that have been, in my view, reasonably well validated. That is, there is a moderate literature showing that claims data are not so bad for studying acute myocardial infarction. They have sensitivities in the 90s and positive predictive values in the 90s.

So, they are not perfect and I think we will have to ask as we review the studied can the amount of uncertainty that we know exists in those account for the effects that we see, or could they obliterate effects that we would like to see and which aren't there.

My sense is that that is probably not a sufficient explanation to dismiss the studies that we are looking at. The issue of bias is one that I think always has to live as a sub-text, but quite frankly, in the studies that do outcomes in the way we have been describing, I don't think that is a serious problem.

For cohort studies, we have to ask are we studying the right population, and here I think we really do have to stop and ask carefully. One is: Are these people selected from the population under study? I think in most of these examples, they are reasonably representative, that is, a study of the

people of Ontario or members of a large health plan.

I think that the data systems that are used to identify the individuals in the cohort are good enough to give us reasonable belief that we are identifying either all the people or a representative sample of them.

I think there is a fair question of whether they are representative of the larger population. We could ask are health plan members systematically different from the general population of individuals who are taking these medications.

The range of studies we have include health plan members. I think that there is reasonable information that they probably are representative, at least with respect to the drug myocardial infarction outcomes that are studied. Studies in Medicare and population-based studies, such as those in Canada, I think also give us reason to think that they are representative.

But there is an important consideration about whether there are issues about the way clinicians practice in those setting that might have a serious impact on selecting individuals. In particular, to the extent that formularies are restrictive of, say, newer or more expensive drugs like the COX-2 inhibitors, but I think we have to ask very carefully whether the factors that would influence the prescribing of one class of drugs over another is likely to seriously impact the risk of these outcomes.

Additionally, if there are cost differentials for these drugs, it may be that there is some form of self-selection that causes individuals who are sicker to

receive these drugs, and I think that it is incumbent on us to expect that to be a problem in every one of these observational studies and to ask how well do these studies do in adjusting for that. I will circle back to that in a moment.

I think we have to be concerned about whether we are studying people who have had prior NSAID exposure, in which case we would be worried about survivor biases, of finding the individuals who are relatively immune to these problems.

Finally, there are study design issues about whether there are restrictions of eligibility that might importantly color the data. For instance, at least one of the studies we are looking at requires individuals to have received at least two dispensings of a nonsteroidal agent in order to be eligible.

That means that you have to live long enough to have two dispensings, so it certainly doesn't tell us anything about the early effects of these drugs, and it might in an important way color the results with regard to later exposure.

There is an important question which is not unique to the observational studies, which is who are the right comparators. We had a number of discussions about that yesterday. I think that all the issues that we discuss with regard to the clinical trials are applicable here. In particular, there is a lot of reason to want to compare to other nonsteroidal users because that gives the best chance of having a group that is similar with regard to underlying disease status and presumably risk of myocardial infarction.

Similarly, it is possible to say that if you really care about COX-2 selective agents, you should compare one COX-2 selective agent to another.

That leaves us in the uncomfortable situation of not knowing what is the risk compared to no use at all, so we have some comparisons that do look at non-users or at least remote users, and that has its strengths. It has the big weakness, of course, of putting us at risk of making comparisons against groups that are unrelated.

So, we are really talking here of mostly about a study like the Kimmel study, not the nested case-control study. The other kinds of concerns that raise red flags are the real concern about losing cases who make the group who are studied unrepresentative.

I would point out to you, for instance, that in the Kimmel study, only half of the MI survivors who were identified were actually interviewed and therefore part of the formal analysis.

We already talked about the fact that since that study was limited to MI survivors, that restricts us to a less serious set of outcomes.

The other problem that really bedevils conventional case-control studies is knowing whether the group of people who are selected as comparators are really comparable.

I think that is one of the reasons that there is so much interest in doing nested case control studies, because at the end of the day it is really extremely difficult

to satisfy oneself that controls really are appropriate.

Much of what we need to be concerned about in these studies is understanding exposures. Part of the issue is understanding how to characterize exposure. This is both a strength and a weakness of these studies.

You will remember I made the point at the outset that if we want to understand how drugs work in actual practice, that we have to do observational studies. On the other hand, that means we have to find a reasonable way to characterize these drugs.

We talked yesterday I think about all the important issues of understanding whether we had to look at absolute dose or cumulative effects or whether the effects start early or whether they start late.

I think that the best of the studies that we are looking at tackle a number of these issues. I will mention in a minute some of the ways that these studies have gone about that.

I think in terms of ascertaining exposure, it is probably reasonable to put the most reliance on the studies that use administrative databases of pharmacy dispensing, but I will just make the point that we have to be clear that these studies are done in situations where we have reason to expect that the administrative databases are correct.

I think all the studies we are reviewing are ones where the investigators were careful to know that the individuals really had a drug benefit that was operating at the moment, that would

likely find the prescription drug exposures that we care about, but as a general proposition, you can't assume that that is the case.

Most health plans have some kind of restrictions on benefits that might lead individuals to change their benefit status, so there would be periods of time when we might know that they had an MI, and we might not know that their drug exposure is at the moment.

I will return to a point that we touched on yesterday, which is that although almost all of the studies that we are talking about report their results as relative risks, a 2-fold increase in risk, a 70 percent decrease in risk. What we really care about is the absolute difference in risk.

So, that is not different between observational studies and randomized studies, but I think it is really a critical piece of our thinking about the problem that we are dealing with.

The second thing that is just worth recalling is that when we talk about a 95 percent confidence interval, that our expectation about where the true value lies is not uniformly distributed over that interval.

Our best guess about where the true value lies is around the point estimate, and if that point estimate is wrong, the large majority of the uncertainty is pretty close to that point estimate, so that it is particularly not helpful, in my view, to pay enormous attention to p values.

The difference between a p value of 0.05, as shown here, and a p value of 0.01 and a p value of 0.13 is not all that

enormous in terms of the biological impact.

I think one of the things that is a particular concern that we need to pay attention to in these studies is the fact that it is easy to look at a lot of different comparisons, and to the extent that we do that, we are going to have to just be careful to know that the strength of any one comparison is weaker than it appears to be.

For instance, this is a quote from one of the studies that we are looking at. We undertook an observational study examining the association between rofecoxib, celecoxib, other non-steroidals and myocardial infarction.

Well, there is no primary hypothesis there, and the results for all of the non-steroidals. They are all interesting to look at, they are all associated with p values. Those p values are all relatively too extreme given the fact that there are so many comparisons.

It is a problem for randomized trials. We talked about subgroup analyses. It is important to do those studies, those subgroup analyses, but absent having specified a principal hypothesis at the outset, I think that we have difficulties in knowing how much weight to put on any particular one.

We talked a lot about confounding. That is one of the most important concerns in randomized trials. I know you all know what confounding is. It wasn't obvious to me when I was making these slides that everyone knew that, but the example, so that we have it in mind is if what we know is drug A versus drug B, and MI or no MI, and we don't take into account

important confounders, we can get importantly incorrect results.

So, here is an example of an aggregate analysis with a relative risk of 1.5 among 2,000 people who are exposed to two drugs. If you break it apart and see that in the high-risk group, drug A accounted for 80 percent of the exposure, and in the low-risk group, drug B accounted for 80 percent of the exposure, you see that in each of those two categories, the high-risk group and the low-risk group, that, in fact, there is no association between drug and outcome, but you have to take them apart to do that.

Well, the good news is if you know what the confounders are, and you have measured them accurately, it is possible to adjust for them, and all of the studies we are looking at do a pretty job of adjusting for the confounders that we know about, so I guess one of the questions is how well do they do at identifying the important confounders.

I would say not bad on a lot of that. That is, if you take, for example, the Graham study or the studies that Wayne Ray did in Tennessee Medicaid, there are a number of strengths. I will sort of stop and back up on the things that make these look like relatively more credible studies in the scheme of the factors that we care about.

They are inception cohorts of nonsteroidal users, that is, they are individuals who had to have been members of the health plan for at least a year before they received their nonsteroidal.

There was a lot of information about their underlying medical status that was

available to the investigators using both claims data and medical record data to ascertain cardiovascular disease along a number of dimensions, utilization of procedures like surgery or angioplasty or diagnostic procedures that are intended to find cardiovascular disease, hospitalizations, emergency room visits, and a substantial amount of information about the medications that these individuals took that was related to or plausibly related to cardiovascular risk factors.

Those large number of factors were used to create separate risk models using only the unexposed, and then to use those risk models to create risk indexes for the individuals to use as an adjuster for underlying cardiovascular risk.

Is it perfect? No. Is it pretty good? It seems to me that it meets the sniff test of saying that it has a reasonable chance of identifying important confounding.

Unfortunately, there are a number of important confounders for which health care systems typically don't have good data, like smoking, OTC NSAID use, obesity, family history, and those are typically much more problematic.

Some of these studies have worked pretty hard to try to either deal with it or understand whether it could be an important problem. One of the handouts we had, for instance, was the study by Schneeweiss and colleagues who looked back at one of the studies by Solomon that was performed in the Medicare data set, and asked how important could these unmeasured confounders be.

They actually had access to information from the Medicare Beneficiary Survey

that asked representative Medicare beneficiaries detailed questions about many of the things that we would ask about. They weren't the people who were involved in that case-control study, but if you assume that the beneficiary survey, members were representative and they gave plausible answers, it is possible to extrapolate back to the source population, and the take-home message from that work, the answer didn't change very much, which is really what we want to know, not sort of the absolute difference, but whether those unmeasured confounders are important enough that they could cause a difference.

I think we still have to be concerned at the end of the day, we still have to be concerned about residual confounding as a potentially important problem.

One way I think that we can draw relative assurance from that work of adjusting for confounding is to ask how much did the estimate of risk change between the unadjusted and the adjusted result.

I think there is a world of difference between an unadjusted result of 10 and an adjusted result of 1.5, and having an unadjusted result of 1.6 and an adjusted result of 1.5. The former, I think the reasonable assumption is we arguably haven't been able to deal with confounding in a way that would let us believe that 1.5 means something.

I think there is a much stronger case to be made when adjusting for important confounders that we know about doesn't change the risk estimate very much, that that is a relative more credible answer.

Having said that, I think that observational studies are best at finding relative risks that are more than 2. I think that I would pay some attention to relative risks of 1.5. I get very nervous about adjusted relative risks of 1.2.

That doesn't mean that they are not right and I don't ignore them, but if we ask is that for sure the answer, my response to that is I am just less certain about that.

I think we are always left at the end, while we spend a lot of time thinking about and adjusting for confounding, and I think we can do a pretty good job of that, it is much harder to adjust for misclassification, and it is essentially impossible to adjust for bias.

So, I think one of the things we have to ask about is are there plausible sources of misclassification and bias, and if there are, in which direction do they work and would they seriously change our interpretation.

We talked about the fact that absolute differences are the important ones that we care about. We have already started to look at data that talks about person level risk and population level risk, so beyond saying that at the end of the day, I think these are the answers that we really need to talk about, not about relative risk.

Personally, I think that we need two kinds of answers. One is what is the information that patients and their physicians need to have to make decisions for them personally about whether to accept certain kinds of treatments in exchange for certain kinds of anticipated benefits.

I think there is a population level concern that we have to have that emerges from the same set of analyses, but takes on a different form.

So, you will be pleased to know that I am wrapping it up now, and I would say that both the cohort and nested case-control designs, which are the bulk of the observational studies that we are looking at, are relatively strong ones and I think deserve the committee's real attention.

I am sorry that not every one of these studies prespecified a primary hypothesis that we can attend to, but we should whenever possible do that. Even though we don't find important effects in some of these studies, I think it is important to recognize that they don't exclude one.

As I have said, I am least certain about attaching great weight to relatively small excess risks even understanding that when they are extrapolated to a large population, they could account for very important public health problems.

Finally, I would say that the things that support the studies' conclusions are the fact that when we do subgroup analyses and look for dose-response effects, that they strengthen the cause-effect relationship, and I think that there is reason to look for consistency across studies.

I take the point that was made yesterday that it is possible that a dozen studies of naproxen could all have the same underlying bias that shift the point estimate in the same direction, but it is not so clear to me what that bias is.

So, I think that we would have to have a reasonable idea of what might explain consistent differences across studies and ask if they are of sufficient magnitude to explain that. As I say, I am not clear that there are those kinds of biases.

I think we have to be cautious about the fact that residual confounding bias and misclassification are all issues with these studies. So, I think that while they add to our discussion, they have to be considered in light of the fact that they are imperfect vehicles.

Thanks. (Applause.)

# Presentation Slides

Note: Some of the slides presented may not be included below.



Interpreting observational studies of cardiovascular risk of NSAIDs.

Richard Platt, MD, MS

Harvard Medical School and  
Harvard Pilgrim Health Care

HMO Research Network Center for Education  
and Research on Therapeutics (CLRT)

February 17, 2005

**Why perform observational studies?**

- Understand experiences of actual users under conditions of actual use – nearly always different from clinical trials.
- Provide timely information by assessing accumulated experience.
- Assess very large populations.

**Types of observational studies**

- Spontaneous reports
- Case series
- Case-control studies – undefined source populations
- Nested case-control studies – well defined source populations
- Cohort studies – retrospective
- Cohort studies – prospective

**Cohort studies: design**

- Identify drug exposed and unexposed
- Assess subsequent outcomes.

**Cohort studies: strengths/weaknesses**

- Some strengths relative to case-control:
  - Better opportunity to select representative exposed and unexposed.
  - Exposure assessment may be less biased.
- Some weaknesses:
  - Exposure status may change over time.
  - Loss to followup.

**Case-control studies: design**

- Identify cases (outcome has occurred) and non-cases (hasn't occurred).
- Assess prior exposures.

**Case-control studies: strengths/weaknesses**

- Some strengths relative to cohort:
  - Efficient – study only cases and a moderate number of controls.
  - Individuals' exposure status can be classified.
- Some weaknesses:
  - Cases/controls may not be representative.
  - Knowing the outcome may bias the exposure ascertainment.

**Nested case-control studies**

- Cases and controls come from a well-defined population.
- Combine many of the strengths of retrospective cohort and case-control studies.

**Observational vs randomized studies: Differences**

- Randomized:
  - Treated/untreated groups more likely to be comparable;
  - Treatment regimen and outcome assessment more certain;
  - Risk factor, adherence info often better.
- Observational:
  - Subjects often more representative;
  - Usage conditions usually more typical;
  - Larger size/ longer duration possibilities permit observation of rare / delayed outcomes.

## Outcomes

- Are the outcomes the right ones?
  - Hospitalized MI (all, survivors),
  - MI+sudden death,
  - Composite thromboembolic.
- Are they measured accurately?
  - Misclassification - claims alone have ~90% predictive value for MI.
  - Bias - no glaring source in studies under review here.

## Subjects: Cohort studies

- Representative exposed subjects
  - Are they representative of the population under study?
  - Are they representative of the larger population?
    - Restrictive formularies or cost barriers may result in risk channeling.
  - May be survivors of prior NSAID courses
  - Eligibility restrictions,
    - Requiring multiple dispensings eliminates those with early MI
- Comparable unexposed subjects
  - NSAIDs? Which ones? Remote users? Never exposed?

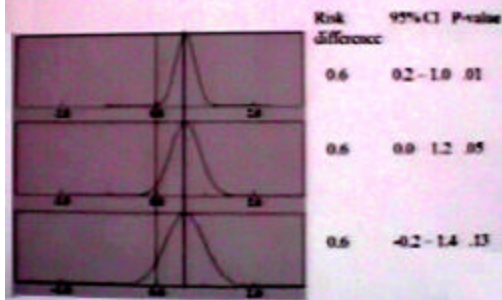
## Subjects: Case-control studies

- Representative cases
  - Loss of cases is serious limitation for conventional case-control study.
  - Limiting to MI survivors restricts to less serious events.
  - Not so problematic in nested case-control studies.
- Representative controls
  - Typically very difficult to be sure controls are drawn from same population as cases.

## Exposures

- Appropriate drugs / appropriate comparators
- Assessing exposure
  - Characterizing exposure
    - High/low dose, early effect, cumulative effect, late effect
  - Ascertaining exposure
    - Can't account for intermittent administration, variable daily dose.
    - Personal recall subject to both misclassification and bias (case-control)
    - Claims data subject to misclassification:
      - Claims data are incomplete if benefits are capped.

## Risk estimates, confidence intervals, P-values



## Multiple comparisons

- Examining many hypotheses increases the probability of finding one that appears more unusual than it really is.
- *"We undertook an observational study examining the association between rofecoxib, celecoxib, and other NSAIDs and myocardial infarction..."*

## Example (1): Confounded risk estimate

	Drug A	Drug B	Total
MI	180	120	300
No MI	820	880	1700
	1000	1000	2000

MI Risk(A) = .18, MI Risk(B) = .12,  
Relative Risk = 1.5

## Example (2): adjusted risk estimate

Confounder	Drug A Higher group	Drug B Higher group	Total
MI	180	120	300
No MI	440	580	1020
	620	700	1320

MI Risk(A) = .2, MI Risk(B) = .2,  
Relative Risk = 1.0

Confounder	Drug A Higher group	Drug B Higher group	Total
MI	20	80	100
No MI	180	720	900
	200	800	1000

MI Risk(A) = .1, MI Risk(B) = .1,  
Relative Risk = 1.0

## Adjusted analyses

- Can correct for confounding
  - If information about the confounders is known.
  - Some confounders are often missing, e.g., smoking, OTC NSAIDs, obesity, family history.
- Residual confounding must always be considered.
- More difficult to correct misclassification and bias

## Quantifying drug-associated risk

- Relative difference vs absolute difference
  - 2-fold increase has different impact in low-risk vs high-risk populations.
- Person-level
  - Number of exposed people required for an extra event to occur.
- Population-level
  - Number of extra events among a specified number of exposed.
  - Number of extra events in the entire (US or other) population.

## Putting it together (1)

- Cohort and nested case-control studies are relatively strong designs.
- The primary pre-specified hypothesis carries the greatest weight.
- Absence of significant effect usually doesn't exclude an important one.
- Small excess risks are the most difficult to interpret, even when they are significant.

## Putting it together (2)

- Factors that support a study's conclusion:
  - Consistency in subgroups and dose-response effects strengthen evidence for cause-effect relationship.
  - Consistency across studies.
- Factors that limit credibility:
  - Residual confounding, bias, misclassification - determine whether direction and potential magnitude can explain effect.