HOW TO READ A MEDICAL JOURNAL ARTICLE

Stephen D. Simon, Ph.D.

OVERVIEW

Reading medical research is hard work. I'm not talking about the medical terminology, though that is often quite bad (if I hear the word "emesis" one more time, I'm going to throw up!). The hard part is assessing the strength of the evidence. When you read a journal article, you have to decide if the authors present a case that is persuasive enough to get you to change your practice.

Some evidence is so strong that it stands on its own. Other evidence is weaker and requires support from other studies, from mechanistic arguments, and so forth. Still other evidence is so weak, that you should not consider any changes in your practice until the study is replicated using a more rigorous approach.

WHAT YOU SHOULD LOOK FOR

When you are assessing the quality of the evidence, it's not how the data are analyzed that's important. Far more important is how the data are collected. Don't agonize over whether the researchers should have used a non-parametric test or whether a random effects meta-analysis is appropriate (just to cite two obscure examples). These are important issues and they generate a lot of debate. But in most cases, the use of one statistical analysis or another is unlikely to make a substantial difference in the conclusions.

The more common and more important threat to the validity of the study relates to how the data are collected, not how they are analyzed. After all, if you collect the wrong data, it doesn't matter how fancy the analysis is. This is good news, because you don't need a lot of statistical training or a lot of mathematical sophistication to assess how the data are collected.

I don't want to imply that data analysis is irrelevant. There are good examples of where a better data analysis led to a different conclusion (Vickers 2001, Skegg 2000). Analysis errors are less frequent and less serious, however, than design errors.

In this presentation, I want to show you what to look for and why. Here are five questions you should ask yourself when reading a journal article.

- Was there a good comparison group?
- Was there a plan?
- Who knew what when?
- Who was left out?
- How much did things change?

In this article, I will justify these questions using anecdotal evidence at times and solid empirical research at other times. I will also highlight real research articles and use them as examples.

IMPORTANT DISCLAIMER

This presentation will review several published journal articles. The intent is to gauge how much evidence each article presents in favor of the efficacy of a new therapy. Some articles will provide a greater level of evidence and some will provide a lesser level of evidence. But articles which provide lesser levels of evidence are still valuable and important.

Nothing stated in this presentation about a particular journal article should be construed as a statement about the quality of that article. The very nature of research requires a series of steps from very preliminary and speculative levels of evidence to more definitive levels of evidence.

Furthermore, when I point out limitations in the evidence presented in a journal article, more often than not, the authors of the article delineate these same limitations in their discussion. But in general, you need to be aware of these limitations because not every journal author is going to be open and honest about the limitations of their research.

CHAPTER 1: WAS THERE A GOOD COMPARISON GROUP?

INTRODUCTION

Almost all research involves comparison. Do woman who take Tamoxifen have a lower rate of breast cancer recurrence than women who take a placebo? Do left handed people die at an earlier age than right handed people? Are men with severe vertex balding more likely to develop heart disease than men with no balding?

When you make such a comparison between an exposure/treatment group and a control group, you want it to be a fair comparison. You want the control group to be identical to the exposure/treatment group in all respects, except for the exposure/treatment in question. You want an apples to apples comparison.

To ensure that the researchers made an apples to apples comparison, ask the following three questions:

- Did the authors use randomization?
- Did the authors use matching?
- Did the authors use statistical adjustments?

Case Study: Vitamin C And Cancer

Paul Rosenbaum, in the first chapter of his book, *Observational Studies*, gives a fascinating example of an apples to oranges comparison. Cameron and Pauling published an observational

study of Vitamin C as a treatment for advanced cancer. For each patient, ten matched controls were selected with the same age, gender, cancer site, and histological tumor type. Patients receiving Vitamin C survived four times longer than the controls ($p < 0.0001$).

Cameron and Pauling minimize the lack of randomization. "Even though no formal process of randomization was carried out in the selection of our two groups, we believe that they come close to representing random subpopulations of the population of terminal cancer patients in the Vale of Leven Hospital."

Ten years later, the Mayo Clinic conducted a randomized experiment which showed no statistically significant effect of Vitamin C. Why did the Cameron and Pauling study differ from the Mayo study?

The first limitation of the Cameron and Pauling study was that all of their patients received Vitamin C and were followed prospectively. The control group represented a retrospective chart review. *You should be cautious about any comparison of prospective data to retrospective data.*

But there was a more important issue. The treatment group represented patients newly diagnosed with terminal cancer. The control group was selected from death certificate records. So this was clearly an apples versus oranges comparison. *It doesn't matter how bad the prognosis was for a patient diagnosed with terminal cancer; it can't be as bad as the prognosis of a patient who has a death certificate.*

Surgical Trial Without Controls

There's another story, unfortunately fictional, which also highlights the importance of a good comparison group.

A prominent surgeon came to give a special lecture at the School of Medicine. He expounded about the great advance that he had made in a specific surgical procedure. At the end of the lecture he drew thunderous applause from the audience. At first it seemed like there would be no questions, but then a young student in the front row raised her hand. "*Did you use any controls?*" she asked. The surgeon seemed to be offended by this question. "*Controls?*" he asked. "*Are you suggesting that I should have denied my surgical advance to half of my patients?*" The rest of the audience grew very quiet. But the young woman was not intimidated. "*Yes,*" she said, "*that's exactly what I meant.* "The surgeon grew even angrier at this, slammed his fist on the podium and shouted "*Why, that would have condemned half of my patients to certain death!*" There was silence for a few seconds. Then the entire auditorium burst out in laughter when the young woman asked "*Which half?*"

Covariate Imbalance

If you want to judge how effective a new therapy is, you need a comparison group. The comparison group would be a group of subjects who receive either the standard therapy or, in some cases, no therapy (e.g., a placebo comparison).

The ideal comparison group should be similar in all respects to the new therapy group except for the therapy itself. For example, the two groups should have a similar range of ages and weights and should be composed of roughly the same proportions in gender and race/ethnicity. The groups should be evaluated concurrently.

Sometimes the groups are dissimilar on some important characteristics. This is known as *covariate imbalance*. Covariate imbalance is not an insurmountable problem, but it does make a study less authoritative.

In a yet to be published research study here at Children's Mercy Hospital, pre-term infants were randomized either to a group that received normal bottle feeding while they were in the hospital or to a nasogastric (NG) tube feeding group. The researchers wanted to see if the latter group of infants, because they had not become habituated to bottle feeding, would be more likely to breastfeed after discharge from the hospital.

The randomization was only partially effective at preventing covariate imbalance. The infants had comparable birth weights, gestational ages, and Apgar scores. There were similar proportions of caesarian section and vaginal births in both groups. But the mothers in the NG tube group were older on average than the mothers in the bottle fed group.

Since older mothers are more likely to breast feed than younger mothers, we had to include mother's age in an analysis of covariance model so that the effect of NG tube feeding could be estimated independent of mother's age.

Beware of situations where the two treatment groups are handled differently. An example of this would be the study of women who use oral contraceptives. These women visit a doctor at least every six months to get their prescriptions renewed. If these women are compared to a women who do not use oral contraceptives, then the former group will probably be evaluated by a doctor more frequently. An increase in the prevalence of certain diseases may actually reflect the fact these diseases are diagnosed earlier because of the frequency of hospital visits.

Similarly, if a certain drug is suspected to have certain side effects, the doctor may question more closely those patients who are on that medication, creating a self-fulfilling prophecy.

Concurrent Controls Versus Historical Controls

Sometimes researchers will assign all of the research subjects to the new therapy. The outcomes of these subjects are compared to historical records representing the standard therapy. This type of study is sometimes called a *historical controls study*. The very nature of a historical controls study guarantees that there will be a major discrepancy in timing. Thus, you have to consider any factors that have changed over time that might be related to the outcome. To what extent might these factors affect the outcome differentially?

The one exception is when a disease has close to 100% mortality (Silverman 1998, page 67). In that situation, there is no need for a concurrent control group, since any therapy that is remotely effective can be detected readily.

DID THE AUTHORS USE RANDOMIZATION?

If the authors of the study decided who would get the new therapy and who would get the standard therapy, we have an *experimental design*. When the authors of the study do have this level of control, they will almost always assign patients randomly.

If the patient did the choosing, if the patient's doctor did the choosing, or if the groups were intact prior to the start of the research, then we have an *observational design*. In an observational design, it is impossible to assign patients randomly.

Information from an experimental design is generally considered more authoritative than information from an observational design because the researchers can use randomization. Randomization provides some level of assurance that the two groups are comparable in every way except for the therapy received.

Randomization requires the use of a random device, such as a coin flip or a table of random numbers. Systematic allocation (i.e., alternating between treatments) is not the same as randomization.

The simplest way to randomize is to layout the treatment schedule in a systematic (non-random) fashion, generate a random number for each value in the schedule and then sort the schedule by the random number.

Randomization ensures that both measurable and unmeasurable factors are balanced out across both the standard and the new therapy, assuring a fair comparison. It also guarantees that no conscious or subconscious efforts were used to allocate subjects in a biased way.

Randomization is not always possible or practical. When this is the case, we have to rely on observational data to draw any conclusions. But when randomization is possible, its use makes a research study more authoritative.

Studies without randomization often require either matching or statistical adjustments. While both matching and adjustments can help to some extent with covariate imbalance, these approaches do not work as well as randomization. In particular, some of the covariate imbalance may be due to factors that are difficult to measure. For example, patients may differ on the basis of

- Psychological state
- Severity of disease
- Presence of comorbid conditions

All of these factors can influence the outcome, but if you can't measure them easily, matching or adjustment is not possible.

So, all other things being equal, an experimental design with randomization is more persuasive than an observational design without randomization. Nevertheless, much can be learned from

non-randomized. Almost everything we know about the risks of cigarette smoking came from observational designs (Gail 1996).

Randomized studies do have some weaknesses. These studies typically rely on the use of volunteers in a narrowly defined research setting. Such situations may not be reflective of how a typical patient behaves in a typical health care setting (Sackett 1997). In this particular aspect, a carefully planned observational design may provide a more relevant comparison.

Another problem with randomized designs is the limit to their size and scope. These limits may make it difficult to detect rare but important side effects. An observational approach like post marketing surveillance is more likely to be successful in these situations.

Studies of the potential harm caused by environmental exposures (such as lead based paint, second hand tobacco smoke, or electro-magnetic fields) are often impossible to randomize because of logistical and ethical issues.

These exceptions, however, do not diminish the value of experimental designs. In situations where observational and experimental studies can both be conducted, most researchers will give greater weight to the evidence in an experimental study.

DID THE AUTHORS USE MATCHING?

Matching is the systematic selection, for every subject in the treatment/exposure group, of control subject with similar characteristics. For example, in a study of fetal exposure to cocaine, you might select infants born to a mother who abused cocaine during pregnancy. For every such infant, you would select a infant unexposed to cocaine in utero, but also who had the same sex, race, and socio-economic status.

Matching will prevent covariate imbalance for those variables used in matching. It will also reduce covariate imbalance for any variables closely related to the matching variables. It will not, however, protect against all covariate imbalance, especially for those covariates that are difficult to measure.

Matching often presents difficult logistical issues, because a matching control subject may not always be available. The logistics are especially difficult when there are several matching variables and when the pool of control subjects that you can draw from is not substantially larger than the pool of treatment/exposed subjects.

Matching is usually reserved for those variables that are known to be highly predictive of the outcome measure. In a cancer study, for example, matching is usually done on smoking. Many neonatology studies will match on gestational age.

Matching In A Case Control Design

When you are selecting patients on the basis of disease and looking back at what exposure might have caused the disease, selection of matching control patients (patients without disease) can sometimes be tricky. You need to find a control that is similar to the case, except for the disease of interest. There are several possibilities, but none of them works perfectly.

- If the cases are people hospitalized for disease, you could choose people who are hospitalized for conditions other than the disease.
- You could ask each case to bring a friend with them. Their friend would be likely to be of similar age and socioeconomic status.
- You could recruit controls from undiseased members of the same family.

You also have to be careful about the variable you use to match. If the matching variable is caused by the exposure or is a similar measure of exposure, then you might "over match" the data and remove the effect of the exposure. Marsh et al. discuss an example of a study examining radiation exposure and the risk of leukemia at a nuclear reprocessing plant. In this study there were 37 workers diagnosed with leukemia (cases) and they were matched to four control workers. Each of the four control workers had to work at the same site, have the same gender, have the same job code, be born within two years of the case, and had to be hired within two years of the hire date of the case.

Unfortunately, there was a strong trend between hire date and exposure. Exposures were highest early in the plant's history and declined over time. So both hire date and exposure were measuring the same thing. When the data was matched on hire date, it artifactually controlled the exposure and pretty much ensured that the average radiation exposure would be the same among both the cases and the controls. This led to an estimate of radiation exposure that was actually slightly negative and not statistically significant.

When the data was rematched using all the variables except for hire date, the effect of radiation dose was large and positive and came close to approaching statistical significance.

Matching In A Randomized Design

In some randomized studies, matching will be used as well. Partly, this is a recognition that randomization will not totally remove covariate imbalance, just like a flip of 100 coins will not always result in exactly 50 heads and 50 tails.

More importantly, however, matching in a randomized study will provide extra precision. Matching creates pairs of subjects who will have greater homogeneity and therefore less variability.

The Crossover Design

The *crossover design* represents a special type of matching. In a crossover design, a subject is randomly assigned to a specific treatment order. Some subjects will receive the standard therapy

first, followed by the new therapy (AB). Others will receive the new therapy first, followed by the standard therapy (BA).

Since the same subject receives both treatments, there is no possibility of covariate imbalance.

When therapies are applied in sequence, timing effects are of great concern. *Are the therapies set far apart enough so that the effect of one therapy is unlikely to carryover into the other therapy?* For example, if the two therapies represent different drugs, did the researchers allow enough time so that one drug was fully eliminated from the body before they administered the second drug?

The possibility of *learning and fatigue effects* are also potential problems in a crossover design.

Special problems arise when each subject receives the standard therapy first and then the new therapy (or vice versa). Many factors other than the change in therapy can cause a shift in the health of patients over time. Unless the researchers can point to other evidence that shows stability of the condition over time, information from this type of study is worthless.

Sometimes difficult circumstances (such as a general failure to respond to the standard therapy) will force the use of this type of design. Further discussion of lack of randomization or other issues with crossover designs can be found in Louis (1992).

DID THE AUTHORS USE STATISTICAL ADJUSTMENTS?

Statistical adjustments represent one way of correcting for covariate imbalance. There are several ways to make statistical adjustments.

First, there are regression adjustments. In a study of breastfeeding, there was an imbalance between the two groups in that one group was much older than the other group. From a regression model, we discover that older mothers breastfeed for longer periods of time, on average, than younger mothers. In fact, for each year of age, the duration of breastfeeding increases by 0.25 weeks on average. So we would adjust the difference of the two groups by 0.25 weeks for every year in discrepancy between the average mothers' ages.

Second, there are weighting adjustments. Suppose a group includes 25 males and 75 females, but in population we know that there should be a 50/50 split by gender. We could re-weight the data, so that each male has a weighting factor of 2.0 and each female has a weighting factor of 0.67. This artificially inflates the number of males to 50 and deflates the number of females to 50. A second group might have 40 males and 60 females. For this group, we would use weights of 1.25 and 0.83.

Both of these adjustments are imperfect, especially when the adjustment variable is imperfectly measured. And these adjustments are impossible if the researchers did not/could not measure the covariates.

SUMMARY—WAS THERE A GOOD COMPARISON GROUP?

Did the authors use randomization? Randomization ensures balance among the two therapy groups with respect to both measurable and unmeasurable factors.

Did the authors use matching? Matching ensures comparable groups during the selection process.

Did the authors use statistical adjustments? Regression or weighting makes adjustments after the data are collected.

CHAPTER 2: WAS THERE A PLAN?

INTRODUCTION

The presence of a plan developed before data collection and analysis adds to the quality of a publication.

- Did the research have a narrow focus?
- Did the authors deviate from the plan?

Case Study: Meat Consumption And Childhood Cancer

Studies of the effects of diet on health often have difficulties with multiple endpoints. An example is a 1994 study of the effect of cured and broiled meat consumption on childhood cancer.

This study examined two types of cancer (acute lymphocytic leukemia and brain tumor). The authors examined five types of meat consumption (ham/bacon/sausage, hot dogs, hamburgers, lunch meats, and charcoal broiled foods). Finally, the authors looked at food consumption both of the child and of the mother during pregnancy.

In the analysis, the researchers used a cut-off to compare low meat consumption to high meat consumption. For example, they compare one or more hamburgers consumed per week to less than one per week. In the text, however, they went further and discussed results with a different cut-off, children who ate two or more hamburgers per week compared to children who ate one or less per week.

This study came under a lot of criticism for its scattershot approach to investigation, though it also had its share of defenders. *There's a saying in statistics, "if you torture your data long enough, it will confess to something."* When a research study has a plan with limited

number of precisely defined hypotheses, the results are more persuasive. When the research has no pre-planned hypotheses, then the results should be considered preliminary and exploratory in nature.

DID THE RESEARCH HAVE A NARROW FOCUS?

A good research study has limited objectives that are specified in advance. Failure to limit the scope of a study leads to problems with multiple testing.

When there are a large number of comparisons being made, the study is considered a fishing expedition. Again, "if you torture your data long enough, it will confess to something."

Swaen et al. (2001) provides empirical evidence that specifying a hypothesis prior to data collection reduced the chances of a false positive finding by a factor of three.

Pollex et al. also show a similar finding in a more light hearted research project. They established a statistically significant association between certain astrological signs to be associated with winning the Nobel prize (Geminis were more likely, Leos were less likely). The authors conclude that *"foraging through databases using contrived study designs in the absence of biological mechanistic data sometimes yields spurious results."*

When Is Multiple Testing Likely To Occur?

Multiple testing often occurs when a researcher examines a large number of subgroups or a large number of endpoints (Howel 1994). Multiple testing problems also occur when a study examines multiple side effects.

When multiple tests are done simultaneously within a paper, there is an increase in the overall Type I error. If 100 tests were performed at alpha = 0.05, you would expect that 5 of those tests would be significant, even if there was nothing at all going on. There are statistical adjustments for multiple comparisons, but these are controversial. Significant results from a large number of unplanned comparisons are useful mostly just for setting future research priorities.

Optimal Cut Points And The Problem With Multiple Comparisons

Researchers will often simplify analysis of a continuous outcome measure by dividing that measure into two or more distinct groups on the basis of cut points. For example, a researcher might categorize his/her subjects as high or low blood pressure when they are above or below a certain value.

An abuse of this approach, called the minimum p-value approach, was noted by Altman (1994). Researchers would examine a variety of cut points and select the one that yielded the most favorable statistics.

For example, some researchers have chosen the cut point from among a large number of possible cut points so as the make the difference in survival times between those patients above the cut point and those patients below the cut point as large as possible.

By examining a multiple number of cut points the chance of drawing a false conclusion (Type I Error) is inflated from the traditional 5% value to a value as large as 40%.

There are several objective ways to select a cut point. Perhaps the best way is to select the cut point prior to looking at the data. This would involve the use of medical judgment.

After the data has been collected, there are some neutral ways of selecting a cut point. The simplest is a median split. If you wanted to create a median split for blood pressure, you would combine the blood pressure data from both groups, and select a value so that half of the blood pressures are larger and half are smaller.

Subgroup Analysis

Subgroup comparisons are a special case of multiple testing. Rather than looking at multiple endpoints, a subgroup analysis compares a single endpoint across several different subgroups within the data.

Subgroup comparisons suffer from three problems. First, the subgroup comparison is usually a non-randomized comparison. Second, the subgroup comparison has less precision because the sample size is smaller. Third, the sample size in a study could be swamped by the potential number of possible subgroups that could potentially be examined.

If you find a subgroup that behaves differently, then you need to ask yourself a few questions. Is this a subgroup that I would have studied a priori if I had been more careful during the planning stage? Is there a plausible mechanism to explain why this subgroup behaves differently? Are there other studies that have similar findings for this subgroup?

DID THE AUTHORS DEVIATE FROM THE PLAN?

Not all research is predictable, so deviations from a pre-designed plan are sometimes necessary. Nevertheless, be cautious about any major deviation from the original research protocol. Some examples of deviations from the plan include:

- Investigating end-points other than those originally specified.
- Developing new exclusion criteria after the study has started.

You need to ask yourself if the authors deviated from the protocol in a conscious or subconscious effort to manipulate the results. Did the authors add other end-points in order to salvage a largely negative study? Were new exclusion criteria targeted to keep “troublesome” subjects out? It is impossible, of course, to discern the motives of the researchers. Nevertheless, for any

deviation or modification to the protocol, you can ask whether this change would have made sense to include in the protocol if it had been thought of before data collection began.

An Example Of A Deviation From The Research Plan

An interesting deviation from the research plan occurs in a randomized double blind control trial for the use of selenium supplements (Clark 1996). The study was initiated in 1983 with basal skin carcinoma and squamous skin carcinoma as the primary end points. The researchers also looked for signs of selenium toxicity.

In 1990, funding was obtained to look at additional secondary end points (total mortality, cancer mortality, and incidence of lung, colorectal, and prostate cancers). While it was relatively easy to add extra endpoints in the middle of the study, the authors acknowledged that this represented a deviation from the protocol.

Another deviation from the protocol occurred when the study was terminated early (January 1996). No statistical changes were found in the primary endpoints, nor was any evidence of selenium toxicity found.

Among the secondary endpoints, however, the authors found statistically significant declines in total cancer mortality and lung cancer mortality. The authors also found statistically significant declines in the incidence of prostate cancer, colorectal cancer, lung cancer and total carcinomas. There was also a decline in overall mortality, though it did not achieve statistical significance.

There were no significant changes in the incidence of nine other types of cancer, including breast cancer, bladder cancer, and leukemia.

Because the significant results occurred in areas that were not originally planned for study, the authors acknowledge that any results have to be considered preliminary. Furthermore, it is unclear what impact the early termination of the study had on the statistics. Early termination of a study can cause serious biases, unless specific rules for early termination are established at the start of the study.

Fraudulent Changes In The Protocol

Detecting fraud in a research study is extremely difficult for anyone, but especially difficult for the reader. A thorough peer review provides a limited level of protection from fraud. Hawkey (2001) proposes that journals should see the original protocols for research studies as part of the peer review process. This practice, which has not yet been widely adopted, would provide some level of protection against fraud.

Sometimes a careful review of the numbers in a study can highlight the possibility of fraud. If a study used randomization, for example, watch out if there is an unexpected and unexplained deviation from a 50-50 split between treatment and control.

Replication of research findings is also a good protection against fraud.

Did The Authors Discard Outliers?

You should be skeptical of any study that removes outliers. Inappropriate removal of outliers can seriously bias the study results.

Sometimes the outliers are more interesting than the bulk of the data themselves. You may gain more insight by trying to uncover the cause of an outlying observation than you would by examining the relatively small effects that occur with the rest of the data.

It is generally a bad idea to remove data points on the basis of their data values alone. If an investigation of an outlier leads to a discovery of a typing error or the inclusion of a subject who did not meet the pre-specified inclusion criteria, then correction or removal of the outlier is appropriate.

If there is no such justification, then the best solution is to leave the outlier alone. Another alternative is reporting data analysis results both with and without the outlier.

SUMMARY—WAS THERE A PLAN?

The presence of a plan developed before data collection and analysis adds to the quality of a publication.

Did the research have a narrow focus? A large number of comparisons limits the amount of evidence that you can place on any single conclusion. Results from a limited number of planned comparisons are considered more authoritative.

Did the authors deviate from the plan? While minor deviations are expected, be cautious about major deviations from the research plan, such as developing new exclusion criteria during the course of the study. In particular, removing outliers without a sound scientific reason is dangerous.

CHAPTER 3: WHO KNEW WHAT WHEN?

INTRODUCTION

Knowledge of group membership during the research study collection can cause problems. When possible, the treatment status should be blinded to the patients, anyone who interacts with the patients, anyone who evaluates the patients, or anyone who collects data from the patients. Even when this is not possible, the randomization list should stay be concealed until the patient agrees to participate in the study and is shown to be eligible for the study.

Acupuncture

Acupuncture is an example of a therapy that is difficult to blind. One study of the effect of acupuncture on the prevention of recidivism among alcohol and other drug abusers (Bullock et al. 1989) used a placebo acupuncture that placed needles 5 mm away from the designated acupuncture point.

The use of placebo acupuncture was intended to keep information about the treatment groups hidden from the patients themselves. The patients knew that they were being "needled," but they did not know if the needles were placed correctly or incorrectly. The assumption for this study is that if acupuncture is effective, then correct application of acupuncture should show a greater effect than incorrect application of acupuncture. There is some controversy, however, over this assumption (Nahin and Strauss 2001).

Because of the nature of acupuncture, the acupuncturists were aware of which patients were which, making this only a partially blinded study. A critique of this study (Sampson 1997) pointed out that there were significant interactions between the acupuncturists and the patients, with opportunities for indirect suggestion and nonverbal communication to occur. One indication that subjects became aware of who was in which group was the fact that there was a far greater tendency for control subjects to drop out of the study.

DEFINITION OF BLINDING

In an experimental study, *it is desirable (but not always possible) to keep the information about the treatments hidden from the patients and anyone involved with evaluating the patient. This is known as "blinding" or "masking."* Blinding prevents conscious or subconscious biases or expectations from influencing the outcome of the study.

There is always some individual who knows which patients get which treatments, such as the pharmacy that prepares the pills and placebos. This is perfectly fine as long as these individuals do not interact with the patients or evaluate the patients.

There is a bit of ambiguity with respect to who is blinded (Devereaux et al. 2001). For example, a survey of 25 textbooks produced nine different definitions of "double blind." Therefore, you should avoid using these terms and focus instead on which individuals are blinded. If you are evaluating an article, look for evidence of blinding for the following groups:

- The patients themselves.
- Clinicians who have substantial interactions with the patients.
- Anyone who assesses outcomes in these patients.
- Anyone who collects data from these patients.

If only some of the above are unaware of the treatment, then the study is partially blinded.

The Effect Of Blinding On The Patient

Blinding prevents the placebo effect from distorting the research results. The placebo effect is a product of "belief, expectancy, cognitive reinterpretation, and diversion of attention" that can lead to psychological and sometimes physiological improvements in situations where the treatment is known to have no effect, such as sugar pills (Beyerstein 1997).

Johnson (1997) lists three specific situations where the placebo effect is of particular concern: when enthusiasm by the patient or the doctor for the new procedure is strong, when outcomes are based on the patient's self-assessment (e.g. quality of life studies), and when the treatment is primarily for symptoms. The placebo effect is less critical for objective outcomes like survival.

A recent study showed that the placebo effect might be overstated in some contexts (Hrobjartsson and Gotzsche 2001). Some of the effects attributed to the placebo are perhaps caused instead by statistical artifacts like regression to the mean or by the tendency of some conditions to resolve spontaneously.

Even without a placebo effect, blinding would still be important to insure uniform rates of compliance. You want to avoid a situation where a patient thinks "I'm in the placebo arm, so it's not really important whether I show up for my follow-up evaluation."

The Effect Of Blinding On The Investigators

The value of blinding also extends to the research team, and should include anyone who interacts with the patients. In a clinical trial of treatments for multiple sclerosis, a pair of neurologists assessed the outcome of each patient (Noseworthy et al. 1994). One neurologist was blinded to the treatment status and one was unblinded. The unblinded neurologist gave substantially lower ratings to patients in the placebo group, which would have led to falsely concluding that one of the treatments was effective.

Researchers can also influence the outcome through their attitudes and through their differential use of other medications (Schulz et al. 2002).

Those who collect data through an interview might probe harder for some patients if they are not blinded. Gail (1996) describes an observational study where the people asking questions about smoking and other risk factors were unaware of when they were interviewing lung cancer patients or controls. Thus, the interviewers could not subconsciously prod more for smoking information among the lung cancer patients.

When Blinding Is Impossible

Unfortunately, there are many situations where blinding is impossible. For example, if you are comparing oral versus rectal administration of a drug, that's pretty hard to conceal from the patient. In general, observational studies cannot be blinded, because the patient and/or their doctor selects the treatment group.

Surgical procedures are often difficult to completely blind. Nevertheless, Johnson (1997) suggests some partial steps at blinding that prevent some of the biases from creeping in. If two surgical procedures use different types of incisions, identical blood or iodine stained opaque dressings could be used to keep the patients unaware of which operation was performed. Also, although the surgeon cannot be blinded to the difference in surgery, those who evaluate the health of the patient after surgery could be kept unaware of the particular operation, so as to insure that their evaluation of the patient is unbiased.

Even though the placebo may look the same, sometimes the doctor may infer which group a patient belongs to, perhaps through noting a characteristic set of side effects. If you are worried about this, ask the doctors to try to identify which treatment group they believe each patient belonged to. If the percentage of correct guesses is significantly larger than 50%, then the allocation scheme was not sufficiently blinded.

Although unblinded studies are considered less authoritative than blinded studies, you should not use blinding as a surrogate marker for the quality of the research (Schulz et al. 2002). For example, Rupert Sheldrake conducted a survey of various journals and showed that blinding was used in 85% of all parapsychology research. But it would be a mistake to claim, as Dr. Sheldrake does, that "Parapsychologists…have been constantly subjected to intense scrutiny by skeptics, and this has made them more rigorous."

Blinding is just of many factors that combine to indicate a study's rigor and quality.

The Problem With Studies Without Blinding

Two researchers have examined studies with and without blinding. These authors found that studies without blinding show an average bias of 11-17% (Schulz 1996; Colditz 1989). In other words, when an unblinded study was compared to a blinded study, the former study tended to estimate a treatment effect that was (on average) 11% to 17% higher than the latter.

Additional evidence of this problem appears in a meta-analysis of the effect of intermittent sunlight exposure and melanoma (Nelemans 1995). When nine studies without blinding were combined, they showed a odds ratio of 1.84 which was statistically significant (95% confidence interval 1.52 to 2.25). When the seven studies with blinding were combined, they showed a much smaller odds ratio (1.17, 95% confidence interval 0.98 to1.39) which was not statistically significant. This is further evidence that unblinded studies are more likely to show statistical significance than blinded studies.

Concealed Allocation

Another important aspect of research is concealed allocation, which is the concealment of the randomization list from those involved with recruiting subjects. This concealment occurs until after subjects agree to participate and the recruiter determines that the patient is eligible for the study.

It is always possible to conceal the randomization list, even when the treatment itself cannot be blinded. Check out all the exclusion criteria and if the subject qualifies, open a sealed envelope which identifies which group the patient belongs to. So, for example, it is impossible to use blinding when comparing a surgical to a non-surgical technique, but the selection of who gets the surgical technique could be hidden from both the patient and the surgeon until after all the selection and inclusion criteria are applied.

Knowledge of treatment order allows the doctors recruiting patients to consciously or unconsciously influence the composition of the groups. They can do this by applying exclusion criteria differentially or by delaying entry of a certain healthier (or unhealthier) subject so he/she gets into the "desirable" group. Unblinded allocation schemes show an average bias of 30-40% (Schulz 1996).

There are many stories of physicians who have tried and succeeded in recruiting a patient into a preferred group. If the treatment allocation is hidden in sealed envelopes, they can hold it up to a strong light. If the sealed envelopes are not sequentially numbered, they can open several envelopes at once. If the allocation is controlled by a central operator, they can call and ask for the allocation of several patients at once.

When a doctor has an overt preference to enroll a patient into one group over another, it raises ethical issues about equipoise and perhaps the doctor should not be participating in the trial.

Concealed allocation only makes sense for a truly randomized study. For convenience, some researchers will allocate in a systematic (non-random) fashion, such as alternating regularly between the two treatments. This is a bad idea. Systematic allocations allow the doctors to guess which group the next patient is going to be allocated to, leading to the same potential problems described above.

Systematic assignment causes an average bias of 15% (Colditz 1989).

SUMMARY—WHO KNEW WHAT WHEN?

Knowledge of group membership, either before or during the data collection can bias the study. Ask yourself who knew what when.

Ideally information about the treatment should be hidden from the patients themselves, anyone interacting with the patients, anyone evaluating the patients, or anyone collecting data from the patients.

The randomization list should be concealed and the treatment assignment should not be revealed until the patient agrees to participate in the study and the recruiting physician has verified that the patient is eligible for the study.

CHAPTER 4: WHO WAS LEFT OUT?

INTRODUCTION

Research studies often have a narrow focus, but sometimes it can be too narrow. When too many patients are left out, those who remain may not be not representative of the types of patients you will encounter.

When you are trying to figure out who was left out and what impact this has, ask the following questions:

- Who was excluded at the start of the study?
- Who refused to join the study?
- Who dropped out or switched therapies during the study?

Nicotine Patches

The Journal of Pediatrics published a study of adolescent smokers in 1996. The researchers recruited 22 volunteers from five public high schools in the Rochester, MN area for participation in a smoking cessation program involving behavioral counseling, group therapy, and nicotine patches. Researchers measured the number of cigarettes smoked, side effects, and blood levels of nicotine.

The purpose of the research was to evaluate "the safety, tolerance, and efficacy of 22 mg/d nicotine patch therapy in smokers younger than 18 years who were trying to stop smoking." The authors also listed a secondary goal, "to compare blood cotinine levels, nicotine withdrawal scores, and adverse experiences with those of adults obtained in previous patch studies." Cotinine is a metabolite of nicotine and provides a useful objective measure of cigarette smoking. It also allowed the authors to examine whether nicotine toxicity was an issue.

This study did not include major segments of the teenage smoking population. The study included only white subjects because there were too few minority students in the Rochester area. Subjects had to get parental permission, excluding smokers who wished to keep their habit secret from their parents. Subjects were also volunteers, and thus could be considered more motivated to quit than the typical teenage smoker.

The study also had a serious drop out rate. Of the presumably thousands of teenage smokers in the Rochester Minnesota area, only 71 volunteers responded to the initial call for subjects. Of the 71 volunteers, 55% met inclusion criteria. Of the remaining 39, 44% declined to attend the initial meeting. Of the remaining 22, 14% were non-compliant. Of the remaining 18, 39% failed to respond to the one year survey. Only 11 completed the entire study (50% of those who started the study; 28% of those meeting inclusion criteria; 15% of the initial volunteers.)

This study had a serious problem with who was left out. The large number of subjects who did not get into the study or who did not complete the study makes it hard to generalize the findings of this research.

WHO WAS EXCLUDED AT THE START OF THE STUDY?

Researchers, trying to minimize variation, will use exclusion criteria to create more homogenous groups. While minimizing variability is good, too much homogeneity can backfire. It's difficult to extrapolate results from a very tightly controlled and homogenous clinical trial to the variation of patients seen in your practice. Ask yourself the question "How similar are my patients?"

For the study to be useful to us, we want the research subjects to be as similar as possible to the patients we see. Watch out for exclusion criteria that leave out large groups of patients. Also be aware that too many research studies exclude women unnecessarily.

Ask yourself whether the geographic location or the type of health care setting places restrictions on the type of patients seen. Tertiary care centers only see patients that are extremely ill. A study of Midwest hospitals will not have a representative number of Hispanic patients compared to the Southwest.

Exclusion Of Elderly Patients

If you are elderly, pat yourself on the back. Your demographic group drives the healthcare economy. You are, by far, the largest consumers of new medications and new therapies. Yet, far too often, these new medications and new therapies are tested on patients much younger (Bayer 2000).

There's a simple reason for this exclusion. When researchers design their experiments, they want a nice clean sample.

Researchers want patients who are ill with one and only one disease. But with older people, several things will break down at the same time (Schellevis 1993).

Researchers don't want patients who are taking a lot of other medications. But older people take so many different drugs that they often qualify for bulk discounts at Walgreen's.

Finally, researchers want patients who are likely to stay alive for the duration of the research study. But older people are likely to die from conditions unrelated to disease being studied.

Although the reasons for excluding elderly patients are understandable, they are still not justifiable. Research done on younger patients cannot be easily generalized to older patients.

Exclusion Of Women

Several decades ago, there was a large study of aspirin as a primary prevention against heart attacks (Physicians Health Study Research Group 1989). This study recruited over 20 thousand physicians and asked them to take either a small dose of aspirin every day or take a placebo. They had to follow these physicians for five to ten years because they wouldn't cooperate and have heart attacks faster. At the completion of the study, the researchers announced that aspirin was highly successful at preventing heart attacks.

There was one major problem with the research sample, though. Every single one of the physicians studied was male. Not a single female was included in the sample. It's not as though this was a problem only for men. Heart disease kills more women than any other condition.

There are some legitimate concerns when testing drugs that might harm a developing fetus, but you can handle this with careful restrictions to women who are not sexually active and/or who are using an effective form of birth control. In addition, some conditions, such as prostate cancer cannot be tested in women.

There is some dispute over whether gender bias exists, with one study arguing that it still occurs (Ramasubbu 2001) and another arguing that it does not (Meinert 2001). *When exclusion of women does occur, it raises troubling questions and hinders your ability to generalize the results of the research.*

Exclusion Of Children

At the opposite extreme from the elderly are children. This group, sadly, is also left out too often.

Children are not little adults. The liver in a child will process drugs quite differently from the liver of an adult. The nutritional demands of a growing child are quite different than those of a fully grown adult. And if you thought that your children became unpredictable as they went through puberty, try looking at them from a medical perspective!

No one wants to see our children used as guinea pigs, and there are special ethical reviews and safeguards that we must comply with when we study children.

Our failure, however, to study children in a careful controlled setting will end up subjecting all children to a large and uncontrolled experiment with no prospect of learning which treatments are safe for children and which ones are harmful.

Volunteer Bias

Quite often, the only patients we are able to study are those who volunteer to help out. The use of volunteers, however, may exclude important segments of the patient population.

Volunteers may differ from the normal population on several critical factors. Volunteers for a study involving cash payments may come more often from economically challenged environments. If a free health check-up is included, volunteers may come more often from people worried about their health status. Volunteers for lengthy studies are less likely to be employed.

Recruiting controls is especially troublesome in a study that involves a painful procedure. Gustavsson (1997) documents volunteer bias in a study of lumbar puncture to obtain cerebrospinal fluid.

In this study, subjects were asked to submit to a lumbar puncture in order to "examine the associations between personality traits and biochemical variables." Of the 87 subjects, 48 declined to participate. The authors were fortunate enough to have measures of personality on both those who participated in the study and those who did not participate.

Those who participated had scores roughly a half standard deviation higher on impulsiveness. They did not differ on other personality traits such as socialization and detachment.

The large difference in the impulsiveness measurement would obviously cloud any attempt to correlate personality traits and biochemical measurements in spinal fluids among those who volunteered.

Hughes et al. (1997) point out the obvious fact that smokers who participate in smoking cessation studies are different from smokers in the general population.

Volunteers In Survey Study

An aspect of volunteering can occur in survey studies. People who volunteer to return a questionnaire are frequently quite different from those who refuse to fill out the survey. In particular, the non-responders tend to be more apathetic. Return rates for surveys vary by the type of survey, but if less than half of the subjects returned the survey, any results are of very limited value. Again, look for efforts to minimize non-response and/or efforts to characterize the demographics of non-responders.

Stocks and Grunnell (2000) examined general practitioners who routinely failed to return mail surveys. A follow-up telephone call assessed demographic characteristics of this group. They were older, less likely to have post graduate qualifications and were less likely to be involved with a teaching practice.

In 1976, Shere Hite published a study on female sexual attitudes that represented the responses of 3,019 surveys. While that sounds impressive, it was a small fraction of the 100,000 surveys that were sent out.

One can speculate on the characteristics of those who failed to respond, but it is a pretty good bet that many of them felt uncomfortable discussing aspects of their sex lives in a survey format. It's obvious that this tendency alone would tend to affect many of the responses in the survey.

What To Look For In Studies Using Volunteers

Examine the incentives and disincentives for participation. Are any incentives or disincentives related to important prognostic factors?

Were the researchers able to characterize various aspects of those who did not volunteer? How similar were the volunteers and non-volunteers?

Do people volunteer themselves into specific treatment groups? If so, we have an observational study.

Some studies involve the use of volunteers who are subsequently randomized into two groups. If this case, some problems will diminish. Comparison between the two groups will be unbiased, but it may be difficult to generalize to a non-volunteer population.

WHO DROPPED OUT OR SWITCHED THERAPIES DURING THE STUDY?

It is inevitable that some patients will drop out during the study. If the number is more than a few, this is a cause for concern.

Dropouts often have a different prognosis than those who stay. Ignoring the dropouts will often paint a rosier picture of the outcome. Was there any effort (financial inducement, follow-up reminders) made to minimize dropouts? Were the authors able to characterize the demographics of the dropouts?

Non-compliance is a common example of stopping or switching therapies. *Were non-compliant patients excluded? Non-compliance is often associated with poor prognosis. Excluding these patients may also paint a rosier picture of the outcome.* Patients should be analyzed in the groups they were randomized to. This is known as "intention to treat" analysis.

Consider a new surgical therapy which is being compared to a standard non-surgical therapy. Some patients randomized to the surgical therapy might die prior to receiving the therapy. This is the most extreme form of non-compliance. These patients should still be analyzed as part of the surgical therapy group. Otherwise the rapidly dying patients will be excluded from the treatment group, but not from the control group, leading to serious bias.

Note that there is still a place for an analysis that excludes noncompliant patients. Such a study answers the question, "What will happen if I prescribe this drug to a group of patients who all take it as directed?" In other words, it looks at a best-case scenario for the tested drug. An intention to treat study asks the question "What will happen if I prescribe this drug to a group of patients that contains both compliant and noncompliant patients?" This presents a more "real-world" estimate of the efficacy of the drug.

SUMMARY—WHO WAS LEFT OUT?

Exclusion of subjects can make the study biased or less generalizable.

Who was excluded at the start of the study? Excessively strict entry criteria in a research study can make it difficult to extrapolate to the types of patients that you normally see.

Who dropped out during the study? A large number of drop-outs during the course of a research study can bias the final conclusions.

CHAPTER 5: HOW MUCH DID THINGS CHANGE?

INTRODUCTION

It's not enough just to assess statistical significance in a study. You need to also make sure that the difference has a practical impact, that it represented a clinically relevant outcome, and that there were sufficient number of patients to provide reasonable precision.

When you are looking at how much things changed, ask yourself the following questions:

- Did the authors measure the right thing?
- Did the authors measure the outcome well?
- Was the change clinically significant?
- Were there enough subjects?

Case Study: Non-Steroidal Anti-Inflammatory Drugs

A 1987 study of non-steroidal anti-inflammatory drugs (NSAID) showed that patients who took these drugs were 50% more likely to develop upper gastrointestinal (UGI) bleeding. This rate was statistically significant at alpha = 0.05. UGI bleeding, however, was rare in both groups. Only 1 case per thousand person years in the controls, 1.5 in the NSAID group. If you see 100 patients a year, you would have to wait two decades, more or less, in order see one excess event of bleeding, on average.

In this article, the authors were up front about the very small increase in risk. Most authors, however, are so relieved to achieve statistical significance that they forget to consider whether the size of the difference will improve clinical practice.

This is summarized well in the following Gertrude Stein quote: "For a difference to be a difference it has to make a difference."

DID THE AUTHORS MEASURE THE RIGHT THING?

There is a tendency to focus on intermediate measures that are easy to assess, but which may or may not be predictive of more important endpoints. Improvement in forced expiratory volume may not translate into a reduction in asthma attacks. A reduction in abnormal ventricular depolarization may not translate into a reduction in the recurrence of heart attacks. If an intermediate endpoint is used, ask yourself whether there is an adequate link between this endpoint and something that is relevant to your patients.

Consider, for example, a study (Leeson et al. 2001) that showed an association between duration of breast feeding and brachial artery distensibility at 20 to 28 years of age. This is a measure of stiffness, and could be considered a surrogate marker for cardiovascular disease in mid and later life. Such a link is tenuous and the authors themselves as well as an accompanying editorial (Booth 2001) admit that no cause and effect relationship between breast feeding and heart disease.

Typically patients are interested in only three things: morbidity, mortality, and quality of life. They don't care about concentration of homocysteine in their blood, or what their CD4 cell count is. They want to know more fundamental questions like "will I die?" or "will I be able to walk up a flight of stairs unassisted?"

Unvalidated Measures

Jadad and Gagliardi (1998) criticize instruments used to rate web sites for the quality of health information. There were 47 such instruments but only 14 discussed how they were created. None of them included measures of validity, which caused these authors to conclude that "Many incompletely developed instruments to evaluate health information exist on the Internet. It is unclear, however, whether they should exist in the first place, whether they measure what they claim to measure, or whether they lead to more good than harm."

Validity is a loaded word that means different things to different people. A general consensus, though, is that a measure is valid to the extent that it measures the thing that it claims to measure and does not mix in things that are unrelated. There are several ways to measure validity, but most of these involve comparison to an external standard.

Short Term Measures

As noted in the introduction, a good measure of the effectiveness of an intervention for schizophrenia, should wait at least six months from the start of therapy. Unfortunately, the typical study lasted 6 weeks or less.

This is a problem for many studies where budgetary limitations force the researchers to focus on short term outcomes. The problem with this is that it is usually easier to get a short term change, especially with interventions that involve behavioral changes (e.g., weight loss through the use of diet and exercise). It is the long term change, however, which is relevant in most cases.

Other Issues

Be careful that you don't focus solely on the outcomes mentioned in the abstract. There is a tendency to report only in the abstract the outcome measures that were statistically significant, rather than the outcome measures most of interest to health care professionals.

Also always consider whether the researcher provided adequate inspection of side effects.

DID THE AUTHORS MEASURE THE OUTCOME WELL?

Research is messy and difficult, so it is not always possible to obtain careful and precise measurements. To what extent are the measurements imprecise and subjective?

Measurement Error

Measurement error is simply the inability to measure an important variable accurately. Measurement error in the outcome variable does not ordinarily cause bias, but measurement error in factors that can predict the outcome are of serious concern.

There are several ways to assess dietary fat intake. The most accurate (and also the most costly) way is through the use of prospectively recorded food diaries.

Sometimes the cost limitations or the retrospective nature of a research study will require a less accurate assessment of dietary fat, such as through an interview. Shapiro (1997) points out that estimation of dietary fat using interviews tends to correlate poorly with estimation using prospective diaries. This would cast doubt, for example, on retrospective studies that tried to associate dietary fat intake with the risk of breast cancer.

Retrospective Data

Retrospective data are data that is collected by looking backwards in time. We obtain this data by asking subjects to recall events that occurred earlier in their lives. We also get retrospective data when we review medical records, birth certificates, death certificates, or other sources of historical data. In contrast, data collected during the course of the study is known as *prospective data.*

Retrospective data are often inexpensive to collect, but you should be concerned about its accuracy. The ability of a subject to recall information is sometimes affected by which group that they are in.

Women who have experienced miscarriages, for example, are more likely to search for and remember events that they feel might "explain" their miscarriage, much more so than a group of comparable control subjects. This differential level of reporting is known as recall bias.

In addition, historical data are often incomplete and it is sometimes difficult to verify its accuracy. Therefore, retrospective data are considered less authoritative than prospective data.

An Example Of Recall Bias

An interesting review of the research process that helped establish that smoking causes lung cancer can be found in Gail (1996). One aspect of the research process was addressing the issue of recall bias.

Doll (1950) studied the association between tobacco smoking and cancer. They selected 709 patients with lung cancer and an equal number of matched controls. The authors were concerned about the retrospective assessment of smoking among patients in both groups. Would patients with lung cancer exaggerate the amount of smoking? Would the interviewers press harder for information about smoking among the cancer patients?

While it would be impossible to totally rule out recall bias, the authors did examine a third group, patients who were diagnosed with lung cancer and who later found out that they suffered from a different disease (false cases). If recall bias was the sole explanation of the difference in reported smoking, then the group of false cases should have had a similar level of smoking with the lung cancer patients. Instead they reported a lower level of smoking. This helped to rule out the possibility that recall bias alone accounted for the higher reported smoking levels in the lung cancer patients.

Confusing Causes And Effects

Another difficulty with retrospective data is that you may not be able to identify which was the cause and which was the effect. Causes have to occur before and effects have to occur after, but when you examine causes and effects retrospectively, you may end up losing information about timing.

There's an old joke about a statistician who was examining the fire department records, including information about how much damage the fire caused, and how many fire engines responded to the blaze. The statistician noticed a strong relationship between the two variables and concluded that the more fire engines you send, the more damage they cause.

The British Medical Journal highlighted a research study where speech patterns were recorded in two groups of surgeons. The first group had two or more malpractice claims filed against them and the second group had none. There was a large difference between the two groups, with the first group having a dominant tone with less concern for the patient. The news report of this research suggested that "dominance coupled with a lack of anxiety in the voice may imply surgeon indifference and lead a patient to launch a malpractice suit when poor outcomes occur."

One reader, however, pointed out that perhaps "being sued is a brutalising and demoralising experience and that this experience fundamentally changes the attitude of doctors towards their patients."

Measurements Without Established Reliability

Reliability means different things in different fields, but the general concept is that a reliable measurement is one that would stay about the same if it were repeated under similar circumstances. Depending on the context, you would establish reliability differently. For example, one way to establish reliability is to have two people make independent assessments and show a good level of agreement. If you are measuring something that is stable over time, then you could take two measurements on different days or weeks and see how well they agree.

Be especially careful about measurements that have some level of subjectivity. If there is no establishment of reliability for these measures, then you have no assurance that the research is repeatable.

WAS THE CHANGE CLINICALLY SIGNIFICANT?

Research results should be quantifiable. Look for measurements of important outcomes that are free from bias.

Knowing that a new therapy is better is not enough information. You need to quantify how much the new therapy is better. In this respect, confidence intervals are better than p-values. *A p-value tells you whether the new therapy is better. A confidence intervals tells you whether the new therapy is better and by how much.* A confidence interval allows you to balance the size of the improvement against the possibility of greater cost or more side effects. Many journals now require confidence intervals instead of p-values.

Statistical methods are sometimes able to detect differences that are so small as to be meaningless from any practical perspective. This is known as statistical significance without clinical significance. Always put the numbers into the perspective of your practice. Try to estimate how many of the patients you see within a year are likely to perform better under the new therapy.

Murray and Teasdale (2000) and Roberts et al. (2000) debate the clinical relevance of a (theoretical) intervention that helps an additional one person out of 10. Does helping "only" one out of every ten patients justify the extra time or money involved? Does it justify an increase in the risk of side effects?

Assessing clinical significance requires clinical judgment. It also needs to factor in preferences of individual patients. It's not easy, and the authors of the research paper should (but usually don't) provide you with their thoughts on clinical significance.

In some studies, however, clinical significance is not important. When you are trying to see if a certain physiologic mechanism can explain why a new therapy works, you just want to know if the mechanism exists or not.

WERE THERE ENOUGH SUBJECTS?

Every research study, especially negative studies, should justify the sample size chosen. *It is unethical to perform research on humans or animals without first demonstrating that the sample size you have chosen is appropriate.*

Justification of sample size is particularly important for a negative study (one where no difference between the standard and new therapies were found) and in studies assessing the equivalence of two therapies.

How Can You Tell If The Sample Size Is Too Small?

Ideally, the authors should provide justification of the sample size in the paper itself. The justification is considered better if it is made a priori (prior to the start of the data collection). If no justification of sample size (e.g., power calculations) is given, examine the width of the confidence intervals. *Very wide intervals indicate an inadequate sample size.*

There Are Many Examples Of Studies With Inadequate Sample Sizes

A revealing study of inadequate sample size appears in Freiman 1992. In a series of 71 publications appearing between 1960 and 1977, the outcome was either percent mortality, percent complications, or a similar outcome that could be measured as a percentage. The authors examined power, the ability of the study to detect either a moderate improvement (25% relative reduction in the outcome) or a large improvement (50% relative reduction in the outcome). For example, if a study showed a 40% mortality in the controls, then a 30% mortality rate in the treated group would be considered a moderate improvement and a 20% mortality rate would considered a large improvement.

The results of the Freiman study were very disappointing.

Of the 71 papers, 57 had greater than a 50% chance for missing a moderate improvement and 31 had a 50% or greater chance for missing a large improvement.

One wonders why anyone would undertake a study when there is such a high probability for failure. You should never initiate a study unless you know that the chance of missing a reasonable improvement is less than 20%.

Special Issues In A Study Of Equivalency

Some studies attempt to show not that a new therapy is superior to the standard therapy, but that it is equivalent. Showing equivalence requires a very careful assessment of sample size.

An example of an equivalence study is when a drug company tests a generic drug and wishes to show equivalence with the (presumably more expensive) brand name drug.

If we applied the traditional testing approach, the company would have a strong disincentive to design the study with an adequate sample size. A small sample size is more likely to show equivalency under the traditional testing framework.

There are several modifications to the traditional testing framework for equivalency studies. The simplest approach uses confidence interval for the ratio of the outcome under new therapy to the outcome under the standard therapy. If both limits of the confidence interval are reasonably close to 1 (e.g., no less than 0.8 and no more than 1.25) then the two therapies are considered equivalent.

SUMMARY—HOW MUCH DID THINGS CHANGE?

Research results should be quantifiable. Look for measurements of important outcomes that are free from bias.

Was there a quantitative measure of the size of the effect? Look for a confidence interval and compare the size of the effect to what you would expect to see in your practice.

Could other factors account for this effect? Look for differences in demographics between the two groups and ask if these differences could explain the results of the research.

Were any important outcomes forgotten? Research results should focus on endpoints that are of interest to your patients.

CHAPTER 6: SPECIAL GUIDELINES FOR META-ANALYSES

INTRODUCTION

Meta-analysis is the quantitative pooling of data from two or more studies. When you are examining the results of a meta-analysis, you should ask the following questions:

- Were apples combined with oranges? Heterogeneity among studies may make any pooled estimate meaningless.
- Were all of the apples rotten? The quality of a meta-analysis cannot be any better than the quality of the studies it is summarizing.
- Were some apples left on the tree? An incomplete search of the literature can bias the findings of a meta-analysis.
- Did the pile of apples amount to more than just a hill of beans? Make sure that the meta-analysis quantifies the size of the effect in units that you can understand.

Declining Sperm Counts

In 1992, the British Medical Journal published a controversial meta-analysis. This study reviewed 61 papers published from 1938 and 1991 and showed that there was a significant decrease in sperm count and in seminal volume over this period of time. For example, a linear regression model on the pooled data provided an estimated average count of 113 million per ml in 1940 and 66 million per ml in 1990.

Several researchers noted heterogeneity in this meta-analysis, a mixing of apples and oranges. Studies before 1970 were dominated by studies in the United States and particularly studies in New York. Studies after 1970 included many other locations including third world countries. Thus the early studies were United States apples. The later studies were international oranges. There was also substantial variation in collection methods, especially in the extent to which the subjects adhered to a minimum abstinence period.

The original meta-analysis and the criticisms of it highlight both the greatest weakness and the greatest strength of meta-analysis.

Meta-analysis is the quantitative pooling of data from studies with sometimes small and sometimes large disparities. Think of it as a multi-center trial where each center gets to use its own protocol and where some of the centers are left out.

On the other hand, a meta-analysis lays all the cards on the table. Sitting out in the open are all the methods for selecting studies, abstracting information, and combining the findings. Meta-analysis allows objective criticism of these overt methods and even allows replication of the research.

Contrast this to an invited editorial or commentary that provides a subjective summary of a research area. Even when the subjective summary is done well, you cannot effectively replicate the findings. Since a subjective review is a black box, the only way, it seems, to repudiate a subjective summary is to attack the messenger.

WERE APPLES COMBINED WITH ORANGES?

Meta-analyses should not have too broad an inclusion criteria. Including too many studies can lead to problems with “apples-to-oranges” comparisons. For example, when you are studying the effect of cholesterol lowering drugs, it makes no sense to combine a study of patients with recent heart attacks with another study of patients with high cholesterol but no previous heart attacks.

There is a lot of variability in how research is conducted. Even in carefully controlled randomized control trials, researchers have tremendous discretion. Sometimes this discretion creates heterogeneity among studies, making it difficult to combine the studies.

Heterogeneity In The Composition Of The Treatment And Control Groups

- Researchers can differ in the inclusion and exclusion criteria.
- Even if these criteria do not differ, there may still be differences in the baseline levels of health in the patients, due to geographical differences in the patient population.
- The controls could be selected independently, or they could be matched to the treatment group subjects.
- The control subjects could be given no treatment, a placebo, or a standard treatment.
- The treatment could differ, such as differences in dose or timing of a drug.

Heterogeneity In The Design Of The Study

- The length of follow-up for the patients could differ.
- The proportion of patients who drop out could differ as well as the proposed statistical treatment of these dropouts.

Heterogeneity In The Management Of The Patients And In The Outcome

- How comorbid conditions are treated.
- How complications are handled.
- How much discretion the patient's physician has in controlling patient care.

The outcome measure itself could differ. For example, Abramson discusses a meta-analysis of hypertension treatment in the elderly. Some of the studies examined cardiovascular deaths and others examined cardiovascular events. Other studies examined cerebrovascular deaths, cerebrovascular events, cardiac deaths, coronary heart disease deaths, and/or total deaths.

Example Of Heterogeneity

In a meta-analysis looking at dust mite control measures to help asthmatic patients, the studies exhibited heterogeneity across several factors. Six studies examined chemical interventions, thirteen examined physical interventions, and four examined a combination approach. Nine of these trials were crossovers, and in the remaining fourteen, there was a parallel control group. Seven studies had no blinding, three studies had partial blinding, and the remaining thirteen studies used a double blind. In nine studies the average age of the patients was only 9 or 10 years, but nine other studies had an average age of 30 or more. Eleven studies lasted eight weeks or less and five studies lasted a full year.

How To Handle Heterogeneity

Some level of heterogeneity is acceptable. After all, the purpose of research is to generalize results to large groups of patients. Furthermore, demonstrating that a treatment shows consistent results across a variety of conditions strengthens our confidence in that treatment.

Nevertheless, you should be aware of the problems that excessive heterogeneity can cause. Mixing apples and oranges may not be so bad; you get a fruit salad this way. But when

heterogeneity becomes too large, you might end up combining not apples and oranges but apples and onions.

Subgroup Analysis

When there is substantial heterogeneity, you can look and compare subgroups of the studies. In a meta-analysis studying atypical antipsychotics, the dose of the comparison drug (haloperidol or an equivalent) varied substantially. Among those studies where the dose of haloperidol was greater than 12 mg/day, atypical antipsychotics showed advantages in efficacy or tolerability. When the dose was less than or equal to 12 mg/day, the atypical antipsychotics showed no advantages in these areas.

Meta-Regression

You can try to adjust for heterogeneity in a meta-analysis. This would work very similarly to the adjustment for covariates in a regression model. For example, Derry et al. used meta-analysis to see if long term aspirin therapy was associated with problems with gastrointestinal hemorrhage. They identified 24 studies that looked at aspirin as a preventive measure against heart attacks. In each of these studies, the rate of gastrointestinal hemorrhages was recorded for both the aspirin group and the placebo or no treatment group. There was substantial heterogeneity in the dosage of aspirin used in the studies, however, with some studies giving as little as 50 mg/day and some as much as 1500 mg/day.

This was actually good news in a way, because the researchers wanted to see if the risk of gastrointestinal hemorrhage was dependent on the dose of aspirin. A plot of the dose versus the risk showed that there was indeed an increased risk but that this risk seemed to be unrelated to the dosage.

Inclusion Of Very Old Studies

When conducting a systematic review how far back should you look? Do you set your exclusion criteria judging on the amount of literature available, or do you limit your search to, say the last 10 years?

That depends a lot on the topic, don't you think? Anything in the field of neonatology would have to have a very narrow time window because the field has changed so much so rapidly.

Other areas where the practice of medicine has been much more stable could have wider time windows. I've seen several reviews that have covered half a century of studies.

If you do select a wide time window be sure to see if your results are similar if you restrict yourself to just the most recent studies.

Ask yourself if there was a sudden change in technology that makes any comparisons before and after that technology an apples-to-oranges comparison. So, for example, a meta-analysis involving AIDS patients should restrict itself to the years following the use of AZT.

Also, ask yourself if researchers in your area tend to discount any research that is more than X years old. If so, then your meta-analysis would lose credibility among those researchers if it included studies older than X.

Sensitivity Analysis

A good approach to heterogeneity is to include a wide range of studies, but then examine the sensitivity of the results by looking at more narrowly drawn subsets of the studies.

The authors can also weight studies by a quality factor and give greater emphasis to randomized studies, which are less likely to have bias. Second, the authors can perform sensitivity analyses. Would the results change if we changed the entry criteria?

In general, heterogeneity increases uncertainty, but this uncertainty cannot be reflected in the width of the confidence limits in the meta-analysis results. *When there is heterogeneity, the most information may reside not in a single estimate of how effective the treatment is, but in a careful examination of the variation in the treatment under different conditions.*

WERE ALL OF THE APPLES ROTTEN?

The quality of a meta-analysis is constrained by the quality of articles that are used in a meta-analysis. *Meta-analysis cannot correct or compensate for methodologically flawed studies.* In fact, meta-analysis may reinforce or amplify the flaws of the original studies.

Observational Studies In A Meta-Analysis

The use of meta-analysis on observational studies is very controversial. Some experts have argued that the biases inherent in observational studies make a meta-analysis an exercise in mega-silliness. But even those experts who do not take such an extreme viewpoint warn that the current statistical methods for summarizing the results of observational studies may grossly understate the amount of uncertainty in the final result.

Sensitivity analysis may be a useful way of highlighting the uncertainties in a meta-analysis of observational studies. Restricting the meta-analysis to selective subgroups of the data can yield insight into the size and direction of biases in observational studies. For example, the researchers could contrast case-control designs with cohort designs, with the latter expected to show less bias, in general. Or the researchers could compare retrospective studies to prospective studies, where again, the latter is expected to show less bias in general. Another possibility for comparison involve comparing studies by the amount to which measurement error is expected to cause problems. In general, researchers should try to stratify the observational studies by known sources of bias.

Meta-Analyses Of Randomized Trials

Some meta-analyses restrict their attention to randomized trials because these studies are less likely to have problems with bias. In other words, they wish to avoid mixing bad observational apples with good randomized trial apples. Sometimes further restrictions can be made on the basis of partial or full blinding of results or on the proper accounting of dropouts.

Concato et al. evaluated clinical topics where there were publications of both randomized controlled trials and observational studies. In this review, the observational studies produced results quite similar to the randomized studies.

Sensitivity Analysis

Even for randomized trials, sensitivity analysis may help. Researchers can use "quality scores" to rate individual studies and then see what happens when studies are restricted to those of highest quality only.

For example, Lucassen et al. looked at interventions for infant colic. Although substituting soy milk for cows milk appeared to have an effect, this effect disappeared when only studies of high methodological quality were considered.

Quality Scores

Many times, the reporting of a study will be inadequate, and this will make it impossible to assess the quality of a study. There is indeed empirical evidence that incomplete reporting is associated with poor quality. In such a case, a "guilty until proven innocent" approach may make sense. For example, if the authors fail to mention whether their study was blinded, assume that it was not. You might expect that authors are quick to report strengths of their study, but may (perhaps unconsciously) forget to mention their weaknesses. On the other hand, Liberati rated the quality of 63 randomized trials, and found that the quality scores increased by seven points on average on a 100 point scale after talking to the researchers over the telephone. So some small amount of ambiguity may relate to carelessness in reporting rather than quality problems.

Another approach is to look at subgroups of studies of a similar design and see if the results are consistent across subgroups. For example, Etminan et al. examined the risk of Alzheimer's disease in users of non-steroidal anti-inflammatory drugs. They identified six cohort studies which showed a combined relative risk of 0.84 (95% CI 0.54 to 1.05) and three case-control studies which showed a much lower combined relative risk, 0.62 (95% CI 0.45 to 0.82).

Meta-Analysis Of Studies With Small Sample Sizes

Some experts advocate great caution in the assessment of meta-analyses where all of the trials consist of small sample size studies. The effect of publication bias can be far more pronounced here than in situations where some medium and large size trials are included.

WERE SOME APPLES LEFT ON THE TREE?

One of the greatest concerns in a meta-analysis is whether all the relevant studies have been identified. If some studies are missed, this could lead to serious biases.

Intentional Exclusion Of Studies

In any meta-analysis, you have to draw a line somewhere. Studies that fail to meet your criteria will not be included in the results. But this can lead to serious controversy. In a Cochrane Review of mammography, seven studies were identified, but only two were of sufficient quality to be used. The Cochrane Review of these two studies reached a negative conclusion, but would have reached an opposite conclusion if the other five studies were added back in.

Publication Bias

Many important studies are never published; these studies are more likely to be negative (Dickersin 1990). This is known as publication bias. The inclusion of unpublished studies, however, is controversial (Cook 1993).

Publication bias is the tendency on the parts of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings. Much of what has been learned about publication bias comes from the social sciences, less from the field of medicine. In medicine, three studies have provided direct evidence for this bias. Prevention of publication bias is important both from the scientific perspective (complete dissemination of knowledge) and from the perspective of those who combine results from a number of similar studies (meta-analysis). *If treatment decisions are based on the published literature, then the literature must include all available data that is of acceptable quality.* Currently, obtaining information regarding all studies undertaken in a given field is difficult, even impossible. Registration of clinical trials, and perhaps other types of studies, is the direction in which the scientific community should move.

Another aspect of publication bias is that the delay in publication of negative results is likely to be longer than that for positive studies. For example, Stern and Simes 1997 showed that among 130 clinical trials, the median time to publication was 4.7 years among the positive studies and 8.0 years among the negative studies. So a meta-analysis restricted to a certain time window may be more likely to exclude published research that is negative.

Many experts are advocating the registration of trials as a way of avoiding publication bias. If trials are registered prospectively (i.e., prior to data collection and analysis) then they can be included in any appropriate meta-analysis without worry about publication bias.

Duplicate Publication

Duplicate publication is the flip side of the publication bias coin. Studies which are positive are more likely to appear more than once in publication. This is especially problematic for multi-

center trials where an individual centers may publish results specific to their site. Tramer et al. (1997) found 84 studies of the effect of ondansetron on postoperative emesis. Unfortunately, 14 of these studies (17%) were second or even third time publications of the same data set. The duplicate studies had much larger effects and adding the duplicates to the originals produced an overestimation of treatment efficacy of 23%. Tracking down the duplicate publications was quite difficult. More than 90% of the duplicate publications did not cross-reference the other studies. Four pairs of identical trials were published by completely different authors without any common authorship.

The Limitations Of A Medline Search

While a Medline search is the most convenient way to identify published research, it should not be the only source of publications for a meta-analysis. Medline searches cover only 3,000 of some 13,000 medical journals (Halvorsen 1992). *The studies missed by Medline and other databases are more likely to be negative studies.*

Furthermore, these databases may fail to index major journals in the third world that can provide important trials. Egger (1997) cites an interesting example of how *Medline excludes most Indian journals, even though these journals are published in English and India produces a significant amount of medical research.*

Foreign Language Publications

Some meta-analyses restrict their attention to English language publications only. While this may seem like a convenience, in some situations, researchers might tend to publish in an English language journal for those trials which are positive, and publish in a (presumably less prestigious) native language journal for those trials which are negative. Interestingly, some studies have shown that the quality of studies published in other languages is comparable to the quality of studies published in English.

Picking The Low Hanging Fruit

In an informal meta-analysis, you should also worry about the tendency for people to preferentially choose articles that are convenient. For example, there is a natural tendency to rely on articles where the full text is available on the Internet or where the abstract is available for review (Wentz 2002).

How To Avoid Bias From Exclusion Of Publications

Search for studies should involve several bibliographic databases, registries for clinical trials, examination of bibliographies of all articles found, the so-called gray literature (presentation abstracts, dissertations, theses, etc.) and a letter calling for unpublished papers to be sent out to key researchers.

Subjectivity

"Blinding," a common tool in other research areas should also be used in meta-analyses. Blinding prevents the differential application of inclusion/exclusion criteria. *The people deciding whether a paper meets the inclusion/exclusion criteria should be unaware of the authors of that paper and the journal. They should also include or exclude the paper on the basis of the methods section only; they should not see the results section until later.*

There is empirical evidence, however, that blinding does not affect the conclusions of a meta-analysis (Jadad et al. 1996, Berlin et al. 1997). Furthermore, blinding takes substantial time and energy.

Data should be extracted from papers by multiple sources and their level of agreement should be assessed. Researchers have found disagreements even on such fundamental concepts such as whether a study was positive or negative (Glass 1981).

Like any other research project, *an overview or meta-analysis needs a protocol.* Unfortunately, many published meta-analyses do not state whether a protocol was used (Sacks 1992). The protocol should specify: the inclusion/exclusion criteria for studies; a detailed description of the process used to identify studies; and the statistical methods used to combine results. Without a protocol, the meta-analysis research is not reproducible.

Authors have been shown to be biased in the articles that they cite in the bibliographies of their research papers (Gotsche 1987; Ravnskov 1992). This same bias could potentially affect the selection of articles in a meta-analysis.

If the authors do not present objective criteria for the selection of articles in their overview or meta-analysis, then you should be concerned about possible conscious or sub-conscious bias in the selection process.

Researchers should also list all of the articles found in the original search, not just the articles used. This allows others to examine whether the inclusion/exclusion criteria were applied appropriately.

Detecting And Correcting For Publication Bias

Sensitivity analysis is also useful here. If the results from published studies are comparable to the results from unpublished studies, for example, then publication bias is less of a concern. Along the same lines, the authors can estimate the number of undiscovered negative studies that would be required to overturn the results of this meta-analysis.

Publication bias is also more likely to occur for studies with small sample sizes. If the results of a meta-analysis are stratified by the sample sizes in the studies, a shift away from the null hypothesis in the smaller studies would be a warning flag about the possibility of publication bias. Statistical and graphical methods have been proposed to examine this further but you should be cautious, however, because sometimes there are other explanations. For example,

smaller studies may tend to use less rigorous designs and these designs may be associated with exaggerated effects (Sterne et al. 2001).

McManus et al. (1998) highlight the importance of consulting experts in the area. They were trying to identify all publications associated with near patient testing, tests where the results are available without sending materials to a lab. The authors used a search of electronic databases, a survey of experts in the area, and hand searching of specific journals. The electronic databases yielded the most number of publications, 50, but still missed 52 publications found by the other two methods.

DID THE PILE OF APPLES AMOUNT TO MORE THAN JUST A HILL OF BEANS?

It's not enough to know that the overall effect of a therapy is positive. You have to balance the magnitude of the effect versus the added cost and/or the side effects of the new therapy. Unfortunately, most meta-analyses use an effect size (the improvement due to the therapy divided by the standard deviation). The effect size is unitless, allowing the combination of results from studies where slightly different outcomes with slightly different measurement units might have been used.

Vote Counting

Avoid "vote counting" or the tallying of positive versus negative studies. Vote counts ignore the possibility that some studies are negative solely because of their sample size. Abramson (1990) notes, for example, a meta-analysis of parenteral nutrition in cancer patients undergoing chemotherapy. Although each of the seven randomized control trials in the meta-analysis failed to achieve statistical significance, the pooled results were highly significant.

Unitless Measures

When you are examining a continuous outcome measure, you should be sure that the results are presented in interpretable units. A measure of effect size does not help you much because it is unitless and impossible to interpret. Consider a store that is offering a sale and announces boldly "All prices reduced by 0.8 standard deviations!"

One meta-analysis shows how important it is to express measurements in interpretable units. Lumley et al. (2001) studied the effect of smoking cessation programs on the health of the fetus and infant. One of the outcome measures was birth weight, and the study showed that the typical program can improve birth weight by a statistically significant amount. The researchers then quantified the amount: 28g (95% confidence interval 9 to 49).

Keep in mind that this is measuring the effectiveness of the smoking cessation program, and not the effect of smoking cessation directly. Typically, you would have to send about 12 to 16 women to these programs in order to get one extra woman to quit smoking. So the effect seen here reflects, in part, how difficult it is to get people to change their behavior.

Still the small size of the effect is important. If you want to assess the costs and benefits of smoking cessation programs, it helps to know that the impact of the typical smoking cessation program on birth weight is quite small. This provides a useful yardstick for comparison to other prenatal interventions.

SUMMARY—SPECIAL GUIDELINES FOR META-ANALYSES

There are four factors you should consider when evaluating a meta-analysis.

Were apples combined with oranges? A review that combines studies that are narrowly drawn offers greater credibility than a combination of heterogeneous studies.

Were all of the apples rotten? Meta-analysis cannot correct the flaws of the existing research studies and may tend to amplify these flaws.

Were some apples left on the tree? Look for efforts to ensure that all relevant publications were identified and considered in the meta-analysis.

Did the pile of apples amount to more than just a hill of beans? Look for overall estimates in units that are meaningful and interpretable. Avoid relying on unitless quantities like the effect size.