





FOUR NEWSPAPER STORIES

FIGURE 2.1 Short caption.

FIGURE 2.2 Here is a long caption that will form into the shape of a paragraph. Here is a long caption that will form into the shape of a paragraph.

Figures 2-5 reproduce four newspaper stories, each of which uses statistical comparisons to address an important medical question:

1. Walking Women. Observational study of 72,488 nurses. Claims a 30 to 40% reduction in heart attacks for women who walk 3 miles per hour at least 3 hours per week.
2. Balding Men. Observational study of 22,000 male doctors. Claims a 36% increase in heart attacks among men balding at the crown of the head.
3. Magnesium Injections. Interventional study of 2316 emergency room patients. Claims a 25% reduction in near-term deaths among heart attack patients who received a magnesium injection.
4. Secretin and Autism. Randomized clinical trial of 56 autistic children. Claims no benefit from administration of the drug secretin.

Which of these articles command belief? They all seem impressively scientific, backed by precise numbers, rates, risks, and biomedical theories. Gourmet statistical consumers, however, can detect important differences between the four stories, differences that tell us how skeptically we should view the conclusions. To change metaphors, this chapter can be thought of as a short course in statistical jui-jitsu by which you can protect yourself against the onslaughts of media medical reportage. (Self-defense is important for non-medical stories too. The ideas of this chapter concern biases and how to recognize them, and apply just as well to politics, sports, psychology, etc as to medicine.)

In fact all four of these stories are well above the media medical average. They were selected because they provide enough detail to permit careful distinctions. Table 2.1 summarizes several key points of comparison. A good rule of thumb is that if you cannot ascertain most of these points in a statistically-based story then it probably doesn't merit serious attention.

	Walking Women	Balding Men	Magnesium	Secretin and Autism
Sample Size:	72,488	22,000 (1500 events)	2316 (~275 events)	56
Cohort:	yes	yes	yes	yes
Intervention:	no	no	yes	yes
Preselected hypothesis and endpoint	maybe	probably not	yes	yes
Causation or Correlation:	cause	correlation		cause
Scientific plausability:	strong	weak	moderate	moderate
Claimed Result:	30-40% less heart attacks	30% increased risk	25% less deaths	no significant effect
Claimed strength of evidence:	?	?	substantial (sig level .02)	no evidence for secretin (sig level > .05)

TABLE 2.1 Summary of some important characteristics of the four newspaper stories in Figures 2-5.

The next sections give a discussion of Table 2's main items.

2.1 **SAMPLE SIZE?**

The articles are arranged in decreasing order of sample size (number of subjects), from 72488 for the walking nurses down to only 56 patients in the secretin autism study. We will see that there is better reason to believe the secretin result than that for the nurses. Quality is more important than quantity in statistical comparisons, and the randomized clinical trial (RCT) methodology of the secretin study bestows an authoritative stamp of quality on it.

Given equal data quality, bigger sample size is better, but even then sample sizes aren't always what they seem to be. The actual comparison in the magnesium study is between the magnesium and placebo receivers in the subset of patients who died following an initial heart attack. There were only about 275 such deaths, not 2316. Similarly, the balding men study tells us that most of the evidence comes from the 1500 men who had heart attacks, not all 22000 in the study, while the walking women story is mute on this point, a mark against it. Well-run clinical trials are usually set up to yield useful measurements on most of the subjects, and it seems likely that this was possible in the secretin study, though the story would be more convincing if we were told what the measured "end-point" was (an IQ test? a measure of communication? adult interaction?)

2.2 **COHORT STUDY?**

All four stories describe "cohort studies", where the statistical comparisons are made within a fixed, well-defined population. To put it another way, each study carries within itself its own control group: the nurses who don't walk much, the men who aren't balding on top, etc.

"Historical controls" are the most familiar alternative to internal controls. They are often invoked by those who feel that it is unethical to use control groups¹ when human lives are at stake. There are three counters to this well-intentioned criticism of hard-hearted medical researchers:

1. Most new drugs and treatments don't work. If it were known beforehand that they were effective then we wouldn't need to experiment. In practice the Treatment group is often in more danger than the controls.
2. Historical controls are usually misleading. Medical techniques change, health conditions change, and most importantly the population of affected individuals changes.

¹ Controls don't have to be placebos. Often the control group in an RCT receives the current standard treatment while the "treatment" group is given the new experimental medicine.

”SAT scores decline” is a meaningless headline if the pool of test-takers is broadening, unless the researchers have managed to follow an identified cohort over time.

3. ”Ethical” studies in which everyone receives the new treatment are usually unconvincing, except in those rare cases where the results are overwhelmingly good (or bad.) In a real sense such studies are ultimately unethical because they don’t lead to an improved choice of treatment for future patients.

Medical ethicists who concentrate on obtaining the ideal treatment for each patient in a trial underestimate the vagueries of medical innovation, and also the altruistic component of clinical trials, in which very sick people contribute, often heroically, to the wellbeing of the next generation.

2.3 INTERVENTIONAL?

Here is the key distinction between the four studies. The first two, walking women and balding men, were purely observational. No attempt was made to change the nurses’ walking habits or to apply Rogaine to balding heads; the groups were observed and the data recorded. It is much easier to accrue large amounts of data this way, and observational studies tend to be huge. However their sample size comes at a great cost. With out intervention we can never be certain that our comparisons are being fairly made. Maybe the nurses that walk more also smoke less, eat better, lead happier home lives. There are methods for improving comparisons in observational studies, as discussed in section 4, but one is never certain that all of the influential factors have been equalized.

Interventional studies like those in the magnesium and secretin stories work directly to equalize the two groups. Their goal is to balance out extraneous but possibly influential factors such as age and gender in order to guarantee a fair comparison. The obvious approach is to assign incoming subjects in a balanced way, half the young people to each group, half the old, half the men, and so on.

This is an awkward task, especially in studies involving people. Usually human subjects arrive one at a time (rather than being kept in a cage awaiting the researcher’s convenience) so that at any one stage the full distribution of ages isn’t known, nor for the other factors. Balancing becomes exponentially more difficult as the number of factors increases. Most importantly, the investigator probably doesn’t know which extraneous factors need balancing. Maybe it’s short versus tall that’s important rather than young versus old, or in the secretin setting maybe it’s parents’ income.

All of these problems are finessed by the simple tactic developed by R.A. Fisher: randomization. Subjects are assigned to treatment or control by the flip of a fair coin, one flip for each new subject. Traditionally the statistician provides the investigator with a series of

sealed envelopes. When a new patient appears the investigator opens the next envelope, reads "treatment" or "control", and acts accordingly. The assignment is more likely to be done by phone connection to a central computer these days, but it's the same idea. Randomization doesn't guarantee perfect equalization but even experiments as small as 10 or 20 patients usually wind up reasonably well-balanced.

The magnesium story says that half of the patients received treatment and the other half placebo but it doesn't say how the assignments were made. Probably they were randomized, but it is possible that Dr. Woods simply alternated between treatment and control. This kind of "time randomization" puts a tremendous burden on the researcher's scientific impartiality. Perhaps patient number 1017 is a nice youngish man with a large family to support. Is he really scheduled for the salt-water control (which Dr. Woods actually doesn't believe is nearly as good as the magnesium), or perhaps could we interchange him with patient number 1018, very old and bad-tempered, who arrived just a few minutes later...?

Most scientists would never resort to conscious bias, but the unconscious is more difficult to discipline. Carried out relentlessly, randomization neutralizes all forms of bias, conscious, unconscious, and accidental. There is one more tactic statisticians use to enforce even-handed assignment of patients, and also a fair reporting of outcomes: Blinding.

A study is Double-Blinded if neither the researchers nor the subjects know which treatment is being used. (Somewhere in a back room is a statistician who does know the assignments!) In the secretin study double-blindness might have been enforced by using placebo pills that looked, smelled, and tasted like the secretins. In this case each randomization envelope could contain an appropriate bottle of pills. Only interventional studies can be randomized and double-blinded, which accounts for much of their virtue as instruments of comparison.

2.4 **PRE-SPECIFIED HYPOTHESIS AND ENDPOINT?**

Chapter 3 concerns the evaluation of evidence in a comparative trial. We will see that an honest evaluation depends on an important assumption: that the hypothesis of interest is specified before the trial begins, or at least before the data is examined, and likewise for the endpoint event used to keep score (e.g. heart attack in the two observational studies.) Picking and choosing among possible conclusions after the data is observed undermines the strength of a comparison and demands a higher level of skepticism from the reader.

Intervention trials like the magnesium study usually begin with a clear hypothesis of interest ("magnesium reduces death following heart attack") and a fixed endpoint ("whether

or not the patient dies within four weeks".) We would be more skeptical of magnesium's efficacy if the four week death window was chosen after the data was examined, chosen perhaps to maximize the apparent effect.

Observational studies are more prone to "data mining", rummaging through a pile of data looking for a conclusion of interest. There is nothing wrong with data mining, and in fact large data bases like those in first two studies are assembled to encourage successful scientific rummaging, but it does greatly increase the probability of spurious results. We will be able to make this point more explicitly in the next chapter when we quantify the strength of evidence in a comparative trial.

It is a good guess that the balding-man hypothesis involved particularly heavy data mining. Baldness and heart attack won't be most doctors' first candidate for a close connection. Perhaps the researchers examined the 22,000 doctor data base for a list of possible factors, height, weight, job, education, before focussing on crown baldness. If so, and it would have been helpful for the newspaper story to have given us more background, increased statistical vigilance is required.

2.5 CAUSATION OR CORRELATION?

Studies 1, 3, and 4 have causative implications: walking more decreases heart attacks, magnesium injections decrease deaths following heart attacks, and, unfortunately, secretin does not decrease autistic impairment. But not so for study 2. Nobody is suggesting that a hair transplant or Rogaine application will reduce the risk of a heart attack. The implication here is purely Correlational. Presumably there is a deeper genetic factory causing both baldness and heart attacks. Baldness is acting as an alarm bell. Your family doctor can use baldness to predict higher coronary risk and proscribe appropriate lifestyle changes. We will discuss correlation in section 3.

Intervention studies are designed to test causation. The treatment of interest is or isn't given to each subject, randomly, so that a clear difference between the treatment and control groups must be Caused by the treatment. (Chapter 3 discusses the definition of a "clear difference".) We would really like to test each subject with and without the treatment, but this is impossible in situations like that of the magnesium study. Even when it is theoretically possible, as in the secretin setting, it may lead to unfair comparisons: whether the treatment goes first or second² can be crucial to its effect.

² There are "crossover trials" in which each subject acts as his or her own control. In this case the choice of the treatment going first or second is done randomly. Crossovers are typically restricted to less drastic experiments where the subjects can reasonably be expected to return to their pre-experimental state after the first administration of treatment or control.

Randomization is a technique for equalizing differences between unequal experimental units, human beings being particularly unequal.

Causation is inherently problematical in observational studies. There is always a lurking possibility that what looks like a cause, walking more decreasing heart attacks, may only be a correlation. In this sense the walking women story is less believable than the balding men since the latter is making a much less ambitious claim.

2.6 **SCIENTIFIC PLAUSABILITY?**

Plausability is a matter of past results, both experimental and theoretical, and also of some scientific common sense. It isn't easy to picture a close connection between baldness and heart attacks, though the researchers suggest a correlational link via testosterone levels, while there is a respectable literature of studies, admittedly mostly observational, linking exercise with reduced coronary risk.

Randomized clinical trials, and their interpretation as discussed in chapter ??, are neutral on the question of scientific plausability. A division of labor is at work here: the statistician evaluates the strength of evidence Within the experiment; it is up to the individual scientist to combine this information with past results³, and of course plausability plays a crucial role in their combination. A level of evidence strong enough to link asbestos and lung cancer may not do for an ESP experiment. The scope of the claimed results is an important component of plausability. If the nurses study claimed that walking also decreased cancer, infertility, diabetes, and HIV our skepticism meters should start clicking loudly. A measured skepticism, augmented by the kind of analyses we have been discussing, is the right attitude for followers of statistical stories in the media.

2.7 **CLAIMED STRENGTH OF EVIDENCE?**

Most statistical studies include an assessment of their conclusion's strength of evidence, though this usually gets lost in media reports. The magnesium story provides enough data to reconstruct the evidential strength: in the language of the next chapter, the support for magnesium's efficacy is at the "substantial evidence" level, stated more precisely as "significance level .02". (The crucial information for this calculation comes in the last two paragraphs of the story. We will return to this example in Part III.) In the secretin story,

³ A more ambitious statistical theory based on "Bayes theorem" (see Part V) aims to directly combine background information with experimental results. This can be helpful to the scientist but it comes at the cost of making difficult probabalistic assumptions about the experimental situation.

the phrase "Statistically the two groups showed no differences" implies that the evidence for an effect did not even reach the "moderate" significance level .05.

Observational studies, with their huge sample sizes, often yield very small significance levels (smaller is better on the significance scale), seemingly indicating very strong evidential strength for the claimed effects. Given the bias potential in non-intervention studies this can be misleading, and such effects can easily dissipate when put to the test of an RCT. Prudent readers demand more from observational studies. It is common for example to demand at least a two-fold change in a claimed relative risk ratio, in addition to statistical significance. (The balding men story claims a three-fold increase for high-risk patients.) This gets us into a question considered in the next chapter, the relationship between statistical significance and scientific import. The stories for our two observational studies don't discuss the strength of evidence, a mark against them, but if they had we still would have had to assess it more cautiously than with the mercury and secretin trials.

Here is a composite scorecard for the credibility of the four stories, as you might evaluate them if you were a government administrator pondering the allocation of the public health budget:

MAGNESIUM. A convincing case is made for the efficacy of magnesium injections, after an initial heart attack, reducing deaths by about 25%. It would be inexpensive and probably effective to educate emergency room physicians on this technique.

SECRETIN. There is a tinge of advocacy (actually anti-advocacy, but that amounts to the same thing) in this story, which arouses suspicion, but the study appears to have been run with full RCT safeguards. Secretin is probably just one more ineffective autism "cure", despite the case-history arguments of its supporters. Don't spend any more of the health budget on secretin treatments.

BALDING MEN. This study doesn't claim much, only a correlation between baldness and heart attacks. There is a strong suspicion of data-mining hanging in the air, so even the correlation has to be discounted. Not much reason to consider this study further.

WALKING WOMEN. The most problematical of the four studies. An impressive causative effect is suggested: if women (and maybe men too) walked more they would decrease their heart attack risk 30-40%. However the study was purely observational so that the "cause" is just a correlation at this point, one that might easily disappear in a randomized trial. Perhaps some public health money should be spent on an intervention trial of exercise and coronary health, but it is premature to endorse a causative link between the two.