Common Biostatistical Problems
And the Best Practices That Prevent Them

Biostatistics 209
April 17, 2012

Peter Bacchetti

Goal: Provide conceptual and practical dos, don'ts, and guiding principles that help in

- Choosing the most meaningful analyses
- Understanding what results of statistical analyses imply for the issues being studied
- Producing clear and fair presentation and interpretation of results

You may have seen a lot of this before, but review of these key ideas may still be helpful. Because of some unfortunate aspects of the research culture we operate in, remembering to follow these guidelines is surprisingly difficult, even when you understand the principles behind them.

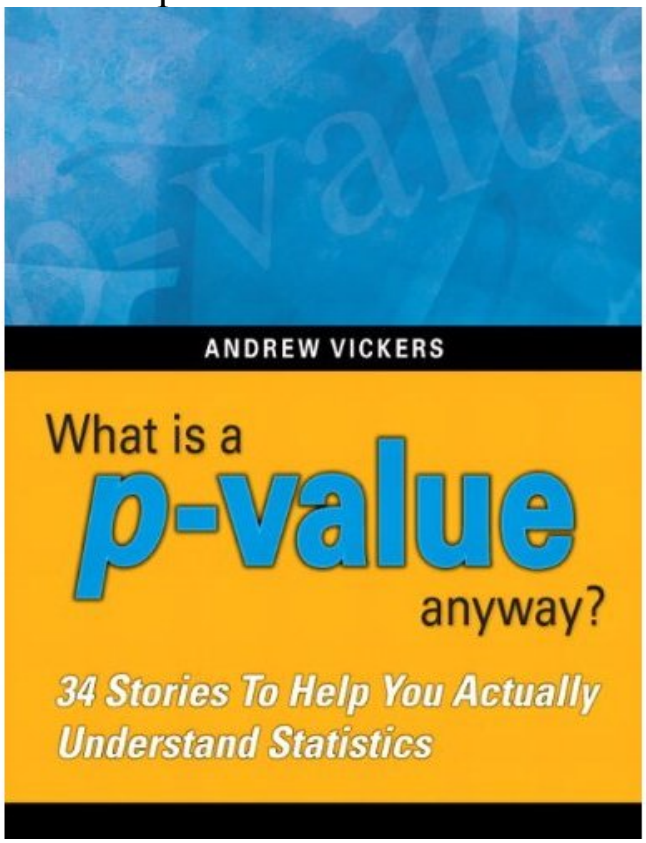
Your class projects will be an opportunity to try these out.

There may be exceptions Rigid, unthinking adherence to supposed "rules" often leads to statistical problems. Please don't consider any of the suggestions provided here to be substitutes for carefully thinking about your specific situation. They should instead prompt more such thinking. See Vickers text chapter 31, The difference between bad statistics and a bacon sandwich: Are there "rules" in statistics?

Please let me know about additions or disagreements

During lecture, or
Later (peter@biostat.ucsf.edu)

Optional reference text



Amazon.com link: http://www.amazon.com/p-value-Stories-Actually-Understand-Statistics/dp/0321629302/ref=sr_1_1?ie=UTF8&s=books&qid=1270360017&sr=8-1

This is a short and very readable textbook that clearly makes many key conceptual points about statistical analysis. This is suitable for complete beginners, but it also can be valuable for more advanced researchers; even faculty statisticians sometimes fail to follow the basic principles that this book explains. The book is not exactly aligned with this lecture, but I think it is useful general reference.
See especially chapters 12, 15, 31, 33.

Despite what you've been taught in this and previous classes about examining estimates with confidence intervals and checking graphical summaries, you may have noticed that p-values are often the primary focus when researchers interpret statistical analyses of their data. Overemphasis, or even exclusive emphasis, on p-values contributes to many problems, including the first, which I also consider to be the biggest problem.

Problem 1. P-values for establishing negative results

This is very common in medical research and can lead to terrible misinterpretations. Unfortunately, investigators tend to believe that p-values are much more useful than they really are, and they misunderstand what they can really tell us.

The P-value Fallacy:

The term "p-value fallacy" has been used to describe rather more subtle misinterpretations of the meaning of p-values than what I have in mind here. For example, some believe that the p-value is the probability that the null hypothesis is true, given the observed data. But much more naïve interpretation of p-values is common.

I hope that no one here would really defend these first two statements:

The p-value tells you whether an observed difference, effect, or association is real or not.

If the result is not statistically significant, that proves there is no difference.

These are too naïve and clearly wrong. We all know that just because a result *could have* arisen by chance alone, that does not mean that it *must have* arisen by chance alone. That would be very bad logic.

But how about this last statement:

If the result is not statistically significant, you have to conclude that there is no difference.

And you certainly can't claim that there is any suggestion of an effect.

This statement may seem a bit more defensible, because it resembles what is sometimes taught about statistical hypothesis testing and "accepting" the null hypothesis. This may seem only fair: you made an attempt and came up short, so you must admit failure.

The problem is that in practice, this has the same operational consequences as the two clearly incorrect statements above. If you are interested in getting at the truth rather than following a notion of "fair play" in a hypothesis testing game, then believing in this will not serve you well. Unfortunately, some reviewers and editors seem to feel that it is very important to enforce such "fair play".

Also see Vickers text Chapter 15, Michael Jordan won't accept the null hypothesis: How to interpret high *p*-values.

How about:

We not only get p>0.05 but we also did a power calculation.

p>0.05 + Power Calculation = No effect This reasoning is very common. The idea is that we tried to ensure that if a difference were present, then we would have been likely to have p<0.05. Because we didn't get p<0.05, we therefore believe that a difference is unlikely to be present. Indeed, you may have been taught that this is why a power calculation is important. But really, this is:

Still no good!

This is still a poor approach, because

Reasoning via p-values and power is convoluted and unreliable.

One problem is that power calculations are usually inaccurate. They rely heavily on assumptions that aren't known in advance. Inaccuracy is theoretically inevitable and empirically verified.

Power calculations are usually inaccurate. A study of RCTs in 4 top medical journals found more than half used assumed SD's off by enough to produce >2-fold difference in sample size.

For example, see a study focused on seemingly best-case scenarios: randomized clinical trials that were reported in 4 top medical journals, *NEJM*, *JAMA*, *Annals of Internal Med*, and *Lancet*:

Vickers AJ. Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology* **56** (2003) 717–720.

Of course, one could do better by re-estimating power after the study is completed. But the assumptions needed for power calculations are still not fully known, and post-hoc power calculations are not considered meaningful. The CONSORT guidelines for reporting randomized clinical trials specifically warn against this practice:

CONSORT guidelines: "There is little merit in a post hoc calculation of statistical power using the results of a trial".

Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG: CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. British Medical Journal 2010, 340:28.
http://www.bmj.com/content/340/bmj.c869.full, bottom of Item 7a.

Why is this not worth doing? Because there is a simpler and better alternative:

Confidence intervals show simply and directly what possibilities are reasonably consistent with the observed data.

Additional references: Use of confidence intervals is widely acknowledged to be superior and sufficient.

1958, D.R. Cox: "Power . . . is quite irrelevant in the actual analysis of data." *Planning of Experiments*. New York: Wiley, page 161.

Tukey JW. Tightening the clinical trial. *Controlled Clinical Trials* 1993; **14**:266-285. Page 281: "power calculations … are essentially meaningless once the experiment has been done."

Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 1994; **121**:200-6.

Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *American Statistician.* 2001;**55**:19-34.

Senn, SJ. Power is indeed irrelevant in interpreting completed studies. *BMJ* 2002; **325**: 1304.

You can see from the phrases such as "Power is quite irrelevant", "the misuse of power", "the fallacy of power calculations", and "power is indeed irrelevant" that there is considerable strength of opinion on this issue. There seems to be a strong consensus.

I've published a critique of conventional sample size planning that discusses these issues, along with many others.
Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. *BMC Medicine*, **8**:17, 2010.
This is also available at: http://www.ctspedia.org/do/view/CTSpedia/SampleSizeFlaws. One of the harms is promotion of Problem 1.

Here are some other situations that make it tempting to believe that a large p-value is conclusive.
How about:

p>0.05 + Large *N* = No effect
p>0.05 + Huge Expense = No effect
p>0.05 + Massive Disappointment = No Effect

Not if contradicted by the CI's! Sometimes we want to believe that a study must be conclusive, because it was such a good attempt or because it looks like it should be conclusive or because nothing as good will ever be done again. But these considerations carry no scientific weight and cannot overrule what is shown by the CI. If the CI is wide enough to leave doubt about the conclusion, then we are stuck with that uncertainty.

Confidence intervals show simply and directly what possibilities are reasonably consistent with the observed data.

Here is an example of the p-value fallacy, based loosely on a class project from many years ago.

A randomized clinical trial concerning a fairly serious condition compares two treatments.

Example: Treatment of an acute infection

The observed results are:

Treatment A: 16 deaths in 100
Treatment B: 8 deaths in 100

And these produce the following analyses:

Odds ratio: 2.2, CI 0.83 to 6.2, p=0.13
Risk difference: 8.0%, CI -0.9% to 16.9%

This was reported as

"No difference in death rates"

presumably based on the p-value of 0.13. This type of interpretation is alarmingly common, but the difference is not zero, which would really be "no difference"; the difference is actually 8%.

Sometimes you instead see reports like these:

"No significant difference in death rates"

This might be intended to simply say that the p-value was not <0.05, but it can easily be read to mean that the study showed that any difference in death rates is too small to be important. Although some journals have the unfortunate stylistic policy that "significant" alone refers to statistical significance, the word has a well-established non-technical meaning, and using it in this way promotes misinterpretation. Certainly, the difference was "significant" to the estimated 8 additional people who died with treatment A.

"No statistical difference in death rates"

This is a newer term that also seems to mean that the observed difference could easily have occurred by chance. I don't like this term, because it seems to give the impression that some sort of statistical magic has determined that the observed actual difference is not real. This is exactly the misinterpretation that we want to avoid. Also see Vickers text chapter 12, Statistical Ties, and Why You Shouldn't Wear One.

A sensible interpretation would be:

"Our study suggests an important benefit of Treatment B, but this did not reach statistical significance."

Finding egregious examples of this fallacy in prominent places is all too easy.

NEJM, **354**: 1796-1806, 2006.
Rumbold AR, Crowther CA, Haslam RR, Dekker GA, Robinson JS. Vitamins C and E and the risks of preeclampsia and perinatal complications. *NEJM*, **354**:1796-1806, 2006

This example from *NEJM* is a randomized clinical trial that concluded:

"Supplementation with vitamins C and E during pregnancy does not reduce the risk of preeclampsia in nulliparous women, the risk of intrauterine growth restriction, or the risk of death or other serious outcomes in their infants."

This very definitive conclusion was based on the following results:

Preeclampsia: RR 1.20 (0.82 – 1.75) This certainly suggests that the vitamins are not effective, because the estimate is a 20% *increase* in the outcome. But the CI does include values that would constitute some effectiveness, so the conclusion may be a bit overstated.

Growth restriction: RR 0.87 (0.66 – 1.16) Here, we have a big problem. The point estimate is a 13% reduction in the outcome, so the definitive statement that vitamins do not reduce this outcome is contradicted by the study's own data. Vitamins *did* appear to reduce this outcome, and the CI extends to a fairly substantial 34% reduction in risk.

Serious outcomes: RR 0.79 (0.61 – 1.02) The same problem is present here, and even more severe. An observed 21% reduction in the most important outcome has been interpreted as definitive evidence against effectiveness. If we knew that this observed estimate were correct, then vitamin supplementation would probably be worthwhile. In fact, the data in the paper correspond to an estimate of needing to treat 39 women for each serious outcome prevented, a rate that would almost certainly make treatment worthwhile.

A less blatant but even higher-profile example is provided by the report on the

Women's Health Initiative study on fat consumption and breast cancer (Prentice, RL, et al. Low-Fat Dietary Pattern and Risk of Invasive Breast Cancer. *JAMA.* 2006;**295**:629-642)

The picture below from *Newsweek* shows a 12-decker cheeseburger next to the text: "Even diets with only 29% of calories coming from fat didn't reduce the risk of disease." This interpretation was typical of headlines. Deeper in the articles, writers struggled to convey some of the uncertainty about the results, but they were hampered by the poor choice of emphasis and presentation in the original *JAMA* publication.

HEALTH

THE NEW FIGHT OVER FAT

BY JERRY ADLER

IF YOU WERE WONDERING what to make of the definitive eight-year study on dietary fat by the Women's Health Initiative released last week, you're not alone. Even some leading researchers were having trouble figuring out what to say about the study's major conclusion: that a low-fat diet did not significantly reduce disease among nearly 20,000 postmenopausal women, compared with a control group who ate what they wanted.

Was Ross L. Prentice of the Fred Hutchinson Cancer Research Center, one of the authors of the study, sounding slightly defensive when he proclaimed that "women can be confident that cutting back on fat ... *certainly won't hurt* when it comes to maintaining a healthy lifestyle"? (Emphasis added.) Did the food industry waste the billions it spent inventing fat-free cookies?

Well, maybe. The problem, says Dr. Marcia Stefanick of Stanford, who heads the steering committee of the WHI, is that the study was designed back in the early 1990s to test an idea that most researchers were already starting to abandon: that the key to health is the total amount of fat in your diet. Instead, most nutritionists now emphasize controlling calories and eating healthy fats—olive and other unsaturated vegetable oils—while avoiding the bad kinds. So it was no great surprise when The Journal of the American Medical Association reported that researchers had found minor reductions, or none at all, in breast or colon cancer or heart disease among women who cut their fat intake on average to less than 29 percent of total calories. (The control group ate a typical American diet with 35 to 38 percent fat.) Those results "are very consistent with what we've seen" in research over the past decade, says Dr. Walter Willett, the prominent Harvard nutritionist, who calls the craze for low-fat everything a "distraction" from good dietary advice.

And that advice—for both women and men—is just what you've been hearing for the past decade: to avoid trans fats (the partially hydrogenated vegetable oils found in processed foods) and restrict saturated fats from meat and dairy products, while consuming a healthy balance of vegetables, fruits and whole grains. "People should stop thinking low fat is the same as healthy," says Stefanick. "The food industry did a great job of selling that, and people believed them." The other advice from nutritionists hasn't changed, either: to exercise and control total calories to avoid obesity. Exercise is important even apart from its effect on weight: it regulates glucose metabolism (lowering the risk of diabetes) and improves bowel function (which may cut the risk of colon cancer). Obesity appears to cause hormonal changes implicated in breast cancer in postmenopausal women, notes Dr. Michael Thun of the American Cancer Society. In the study, the women who ate a lower-fat diet didn't lose weight, but neither did they gain—a fact that gives small comfort to either side in the great struggle between the authors of low-fat and low-carb diet books.

Even after this definitive study, though, most nutritionists (except for those in the Atkins ultra-low-carb camp) still think there's a benefit to limiting fat consumption. Buried in the larger story of the study was the intriguing statistic that

Even diets with only 29% of calories coming from fat didn't reduce the risk of disease.

Read Dr. Dean Ornish's new column on dieting, nutrition and health

Check out the best Web sites for learning about Black History Month

Read more about the Hazelden drug center at Newsweek.com on MSNBC

PHOTO ILLUSTRATION BY NEWSWEEK. PHOTOS (FROM LEFT): HEMERA TECHNOLOGIES—ALAMY; ELIZABETH WATT—JUPITER IMAGES

FEBRUARY 20, 2006 NEWSWEEK 69

The primary result was an estimated 9% reduction in risk of invasive breast cancer:

Invasive Breast Cancer
HR 0.91 (0.83-1.01), p=0.07

An accurate sound bite would have been, "Lowering fat appears to reduce risk, but study not definitive".

An interesting additional result was:

Breast Cancer Mortality HR 0.77 (0.48-1.22)

The estimate here is a more substantial reduction in risk, but the uncertainty is wider. If this estimate turned out to be true, this would be very important.

Unfortunately, the authors chose to primarily emphasize the fact that the p-value was >0.05. This gave the clear (and incorrect) impression that the evidence favors no benefit of a low-fat diet. The primary conclusion in the abstract was:

From *JAMA* abstract:
"a low-fat dietary pattern did not result in a statistically significant reduction in invasive breast cancer risk"

I believe this emphasis promoted considerable misunderstanding.

Best Practice 1. Provide estimates—with confidence intervals—that directly address the issues of interest.

This is usually important in clinical research because both the direction and the magnitude of any effect are often important. How to follow this best practice will usually be clear, as it was in the above examples. Ideally, this will already have been planned at the beginning of the study. Often, an issue will concern a measure of effect or association, such as a difference in means, an odds ratio, a relative risk, a risk difference, or a hazard ratio. Think of what quantity would best answer the question or address the issue if only you knew it. Then estimate that quantity.

Often followed (but then ignored when interpreting)

The above examples provided estimates and confidence intervals, but then ignored them in their major conclusions, which were based only on the fact that the p-values were >0.05.

BP2. Ensure that major conclusions reflect the estimates and the uncertainty around them.
In particular:
BP2a. Never interpret large p-values as establishing negative conclusions.

This is the practice that is too often neglected, particularly for studies with p>0.05, leading to Problem 1. The estimates and CI's should contribute to the interpretation, not just the p-value.

Here are some basic principles to follow when interpreting your results:

The estimate is the value most supported by the data This means that a conclusion is inappropriate whenever it would be wrong if the estimate turned out to be the true value.

The confidence interval includes values that are not too incompatible with the data This means that conclusions are exaggerating the strength of evidence whenever they imply that some values within the CI are impossible or very unlikely.

The study provides strong evidence against values outside the CI If all important effects are outside the CI, then you can claim a strong negative result.

Here is an example of a strong negative result that is well supported

NEJM, **354**: 1889-1900, 2006
Outcomes among Newborns with Total Serum Bilirubin Levels of 25 mg per Deciliter or More
Thomas B. Newman, M.D., M.P.H., Petra Liljestrand, Ph.D., Rita J. Jeremy, Ph.D., Donna M. Ferriero, M.D., Yvonne W. Wu, M.D., M.P.H., Esther S. Hudes, Ph.D., and Gabriel J. Escobar, M.D. for the Jaundice and Infant Feeding Study Team. N Engl J Med 2006; 354:1889-1900.

Conclusion: "When treated with phototherapy or exchange transfusion, total serum bilirubin levels in the range included in this study were not associated with adverse neurodevelopmental outcomes in infants born at or near term."

This is supported by a statement in the abstract concerning the CI's:
Support: "on most tests, 95 percent confidence intervals excluded a 3-point (0.2 SD) decrease in adjusted scores in the hyperbilirubinemia group."

What if results are less conclusive? Such as with the vitamin study discussed above. For the results below:

Growth restriction: RR 0.87 (0.66 – 1.16)
Serious outcomes: RR 0.79 (0.61 – 1.02)

an honest interpretation of what can be concluded from the results would be something like this:

"Our results suggest that Vitamin C and E supplementation substantially reduces the risk of growth restriction and the risk of death or other serious outcomes in the infant, but confidence intervals were too wide to rule out the possibility of no effect."

This interpretation reflects the key facts that 1) the estimates are big enough protective effects to be important and 2) the uncertainty around them is too large to permit a strong conclusion that any protective effect exists.

What would have happened if the vitamin paper had been submitted with this more reasonable interpretation?

But then the paper probably won't end up in NEJM! Unfortunately, this more accurate interpretation would probably have greatly reduced the paper's chance of acceptance.

The "elephant in the room" when it comes to conflict of interest:

- We are all under pressure to make our papers seem as interesting as possible.

There is careful attention to financial conflicts of interest in medical research, but the conflict between scientifically accurate interpretation versus maximizing interest, getting into a high prestige journal, and generally attracting attention (and citations) is largely unrecognized. It is always present, can have a strong influence on presentation and interpretation, and it gets little attention.

The p-value fallacy can help make "negative" studies seem more conclusive and interesting.

The most prestigious journals tend to prefer results that seem conclusive, perhaps because they are trying to serve clinicians in addition to scientists, and clinicians can make the most use of unambiguous results. Although there is a lot of pressure to make results seem as interesting as possible and to get into such journals, this should only go so far. Using the p-value fallacy to make a study seem definitive in one direction when it is really suggestive in the other direction would clearly be going too far. I doubt that this is often deliberate. In this case, the authors may have felt that $p>0.05$ was definitive because the study was large and expensive, or perhaps because they had done a power calculation (but their assumptions were wildly off, as usual with power calculations).

Be vigilant (and be honest)!

The usual safeguards against bias due to conflict of interest are disclosure and correspondingly increased vigilance. Because this conflict is always present, the only obvious response is to always be vigilant.

There is another Best Practice that is useful for preventing Problem 1:

BP3. Discuss the implications of your findings for what may be true in general. Do not focus on "statistical significance" as if it were an end in itself.

This may seem like a subtle distinction, but it is fundamental. We do research to learn about what is true in general in the real world, and p-values and statistical significance do not characterize the state of nature—they are properties of a particular study. Interpretation should focus clearly on what evidence the study provides about what is generally true, not treat statistical significance as an end in itself. Statistical significance is only important by virtue of what it conveys about the study's evidence. Because of the extreme emphasis on statistical significance in medical research, this is often forgotten and we slip into thinking that statistical significance is what really matters. Statistical significance implies strong evidence for an effect, but this is usually not all that is important, and the implications of lack of statistical significance are much less clear.

In the case of WHI, we care about the biological effect of dietary fat and about actual cases of breast cancer that could be prevented. The disconnect between the author's statements and how they were interpreted illustrates why this Best Practice is important.

WHI conclusion:
"a low-fat dietary pattern did not result in a statistically significant reduction in invasive breast cancer risk … However, the nonsignificant trends … indicate that longer, planned, nonintervention follow-up may yield a more definitive comparison."

Newsweek followup article: The week after the article shown above, *Newsweek* published a followup concerning the difficulties that the press and the public have in understanding scientific results, particularly about diet research. (I would add that scientists also have difficulty with these issues.) Despite this focus, the writers still did not understand what the WHI article stated. I believe that this was because they assumed—quite reasonably, but incorrectly—that the article must be addressing the real-world question.
"The conclusion of the breast-cancer study—that a low-fat diet did not lower risk—was fairly nuanced. It suggested that if the women were followed for a longer time, there might be more of an effect."

Both the major conclusion from the abstract and the caveat that followed it concerned statistical significance rather than what is really true. Although the "nonsignificant trends" were mentioned, their implications for the important issues were not discussed. The *Newsweek* writers mis-translated these into more relevant—but incorrect—statements. The statements in yellow are not the same, and the statements in blue also do not match—the authors meant that the study's results may reach $p<0.05$, not that the difference will get bigger.

Because the WHI authors chose to completely neglect any direct assessment of the implications of their findings for what may really be true, I believe that they made serious misunderstandings virtually inevitable.

BP3 and BP2 are complementary. Following BP2 will usually keep you on track for BP3, and vice versa.

While it may seem easy to understand that the p-value fallacy is not valid, it can be surprisingly hard in practice not to lapse into interpreting large p-values as reliable indications of no effect. In some earlier years, for example, most written projects for this class did contain such lapses.

Easy to slip into relying on "p>" reasoning This may be because

- Yes or No reasoning more natural
- Focus on p-values engrained in research culture As we saw for WHI
- Real level of uncertainty often inconveniently large, which can make results seem less interesting The vitamin study is a good example of this, as discussed above.

To avoid this problem, you therefore need to

Be vigilant

- Double-check all negative interpretations
- Examine estimates, confidence intervals

How to check negative interpretations:

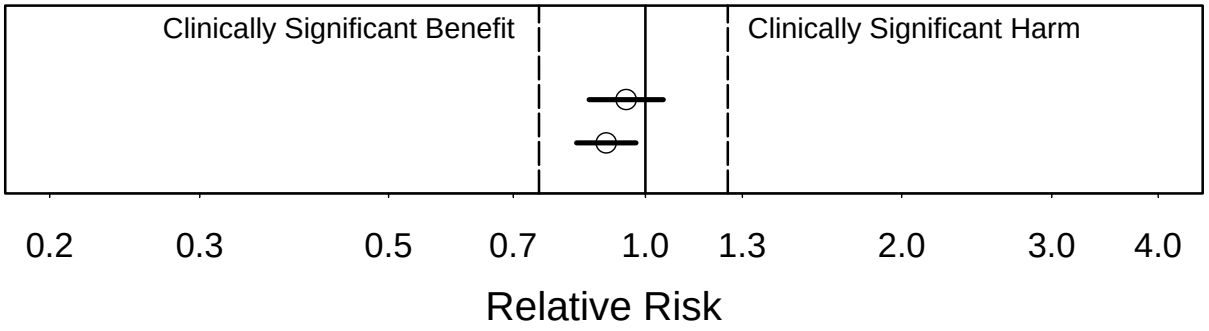
Perform searches for words "no" and "not" Whole word searches on these two terms should find most negative interpretations of statistical analyses.

Check each negative interpretation found and ask yourself

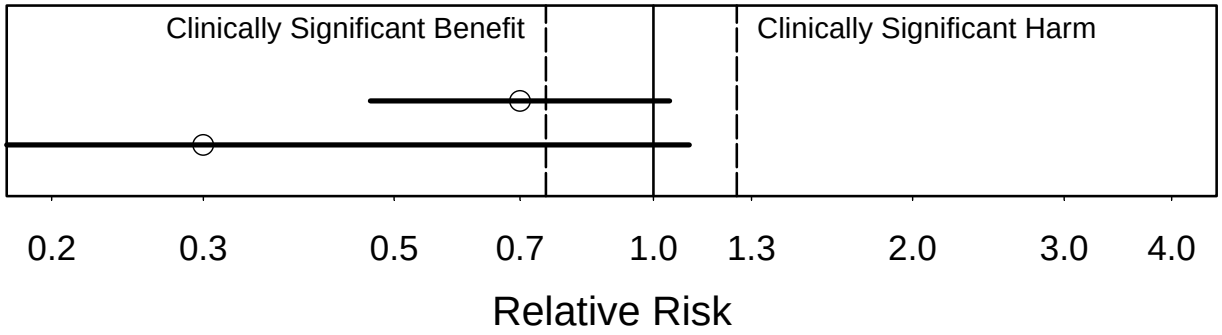
- Is there an estimate and CI supporting this?
- What if the estimate were exactly right? Would the conclusion still make sense?
- What if the upper confidence bound were true? Does the conclusion allow for this possibility?
- What if the lower confidence bound were true? Does the conclusion allow for this possibility?

Additional searches: "failed", "lack", "absence", "disappeared", "only", "rather", "neither", "none" Negative interpretations sometimes use these words, so you can also check them to make sure you didn't miss anything.

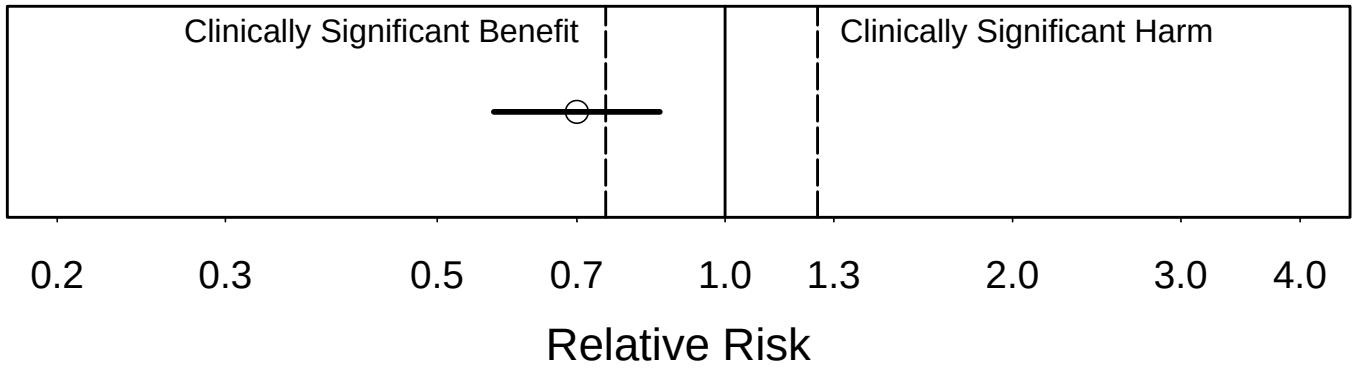
The following figures show some concrete examples of how to interpret estimates and CI's. These assume a somewhat idealized situation where we have exact limits on what is clinically important, but they illustrate the main ideas. Often it will be more practical to first calculate the estimates and CI's and then consider whether the values obtained are large enough to be clinically important. In some cases, it may be hard to argue that any effect, if real, would be too small to be important.
Many detailed examples of how to word interpretations are available at http://ctspedia.org/do/view/CTSpedia/ResultsInterpretation



We found strong evidence against any substantial harm or benefit Because we have strong evidence against any values outside the CI, both these cases argue strongly that any effect is clinically unimportant. Note that this is true even though one is statistically significant.

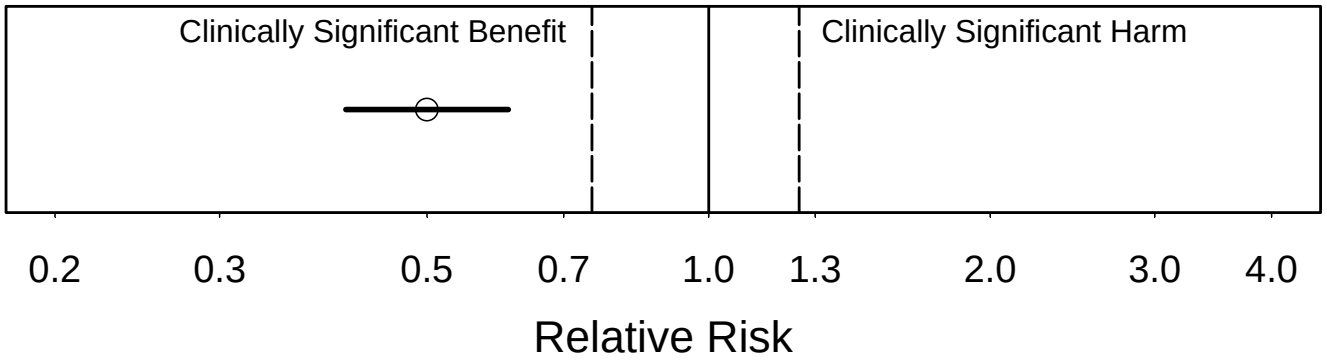


Suggestion of substantial benefit The estimate would be an important benefit if true
May be no effect (not statistically significant) The CI includes no effect
Which of the two results would be more exciting? I think the lower one is, even though it has wider uncertainty, because the estimate is so much better.

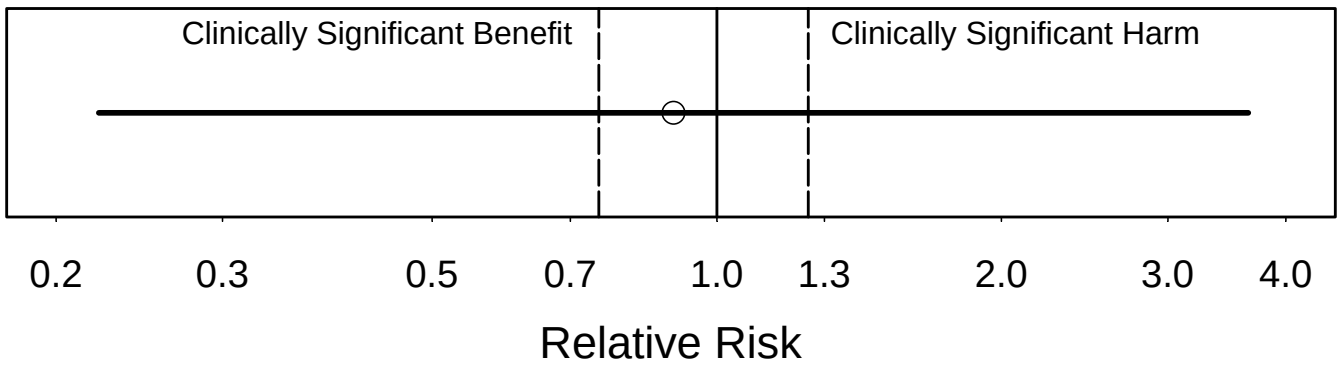


Strong evidence of benefit (statistically significant)

Substantial benefit appears likely, but CI too wide to rule out clinically unimportant benefit The CI includes some benefits that would be too small to be clinically important.



Strong evidence of substantial clinical benefit This is the most satisfying type of result. Even the upper confidence bound is in the substantial benefit range.



No conclusions possible due to very wide CI This is the least satisfying type of result. There is very little information in the study data.

Also see online resource at http://ctspedia.org/do/view/CTSpedia/ResultsInterpretation

Here is an
Example from a typical collaboration:

First draft text:
"There were no statistically significant effects of DHEA on lean body mass, fat mass or bone density."

Final wording:
"Estimated effects of DHEA on lean body mass, fat mass, and bone density were small, but the confidence intervals around them were too wide to rule out effects large enough to be important."

I find that modifications like this are needed in the majority of papers that I am asked to co-author.

We can better understand the limited value of large p-values by noting what they *are* good for.

Are large p-values good for anything? I think yes, but care is needed to recognize such situations and not to overstate conclusions.

"Due diligence" situations where you just want to show that you took some reasonable precautions.

Checking for possible assumption violations when little suspicion is such a due-diligence situation.

Just need to state that you checked and nothing jumped out; don't need to prove that no violation was possible

Be sure to use statements like "no interaction terms of treatment with other predictors in the model had $p<0.1$" rather than "there were no interactions of treatment with other predictors in the model," which would be an instance of the p-value fallacy. Another example is "We checked linearity assumptions by adding quadratic terms for each linear predictor, and none had $p<0.05$", not "there was no non-linearity," which again would be based on the p-value fallacy.

Here is an example from a paper I wrote (Bacchetti P, Tien PC, Seaberg EC, O'Brien TR, Augenbraun MH, Kral AH, Busch MP, Edlin BR. Estimating past hepatitis C infection risk from reported risk factor histories: implications for imputing age of infection and modeling fibrosis progression. *BMC Infectious Diseases*, **7**:145, doi:10.1186/1471-2334-7-145, 2007):

We note that the confidence intervals were not narrow enough to rule out potentially important interactions, but in the absence of strong evidence for such interactions we focus on the simpler models without them.

Problem 2. Misleading and vague phrasing

We failed to detect …
Our results do not support …
We found no evidence for …
Our data did not confirm …

"There is no scientific evidence that BSE [Mad Cow Disease] can be transmitted to humans or that eating beef causes it in humans."

-- British Prime Minister John Major, 1995

Of course, it turned out that BSE *was* transmitted to humans, and over 150 people died from it. "There is no evidence" is commonly used to give the impression that there is evidence on the other side. Although this and similar phrases sound "scientific", they promote sloppy reasoning.

Wording similar to this is popular in politics and advertising, because it gives the misleading impression of a strong case against a conclusion when no such case exists, without being technically incorrect. For the same reasons, the strange popularity of these phrases in scientific writing is disturbing. While such phrases may be technically correct, they are bound to be misread as implying evidence against an association or effect. They often involve or promote the p-value fallacy. And when there *is* strong evidence against an effect, they are too weak.

BP4. State what you did find or learn, not what you didn't.

What a study *did* find is what is interesting, and any conclusions or interpretation should be based on this. Like BP3, this also helps with following BP2.

This prevents deception, but also can make statements clearer and stronger.

Oddly, investigators often understate their conclusions using weak phrasing. The phases above seem like safe, conventional ways to state interpretations, despite their drawbacks. I still find that these phrases sometimes pop into my mind when I think about how to interpret results.

FRAM, nationwide study of fat abnormalities in persons with HIV This was a large study that carefully investigated changes in fat in various anatomical sites among persons with HIV. Its results strongly contradicted established thinking in this area, which was that visceral fat (known as "VAT") increased as peripheral fat decreased and these two changes were causally linked.

Among many results supporting its conclusion of no reciprocal change was the following, showing that peripheral fat loss did not have any substantial association with central fat gain (note upper confidence bound of 1.06).

Peripheral fat loss association with central fat gain, OR: 0.71, CI: 0.47 to 1.06, P = 0.10.

Despite the strong results, some phrasing in an early draft was:

First draft: "our results do not support the existence of a single syndrome with reciprocal findings."

This was revised to read more appropriately:

Final: "We found evidence against any reciprocal increase in VAT in HIV-infected persons with peripheral lipoatrophy" *JAIDS*, **40**:121-131, 2005

Another example:

Safety of cannabinoids in persons with treated HIV

Marijuana effect on $\log_{10}$ VL: -0.06 (-0.26 to 0.13)
Dronabinol: -0.07 (-0.24 to 0.06)

These upper confidence bounds were considered too small to be important, so this was strong evidence against any substantial harm.

First draft: "Overall there was no evidence that cannabinoids increased HIV RNA levels over the 21-day study period."

Final: "This study provides evidence that short-term use of cannabinoids, either oral or smoked, does not substantially elevate viral load in individuals with HIV infection." *Ann Intern Med*, **139**:258-66, 2003

Problem 3. Speculation about low power

This is the opposite of claiming high power to bolster a negative conclusion based only on a p-value, discussed above for problem 1. Sometimes investigators want to argue that their hoped-for results could still be possible, so they mention that power might have been too low and that could be why they didn't see what they expected. This again is too convoluted and unreliable to be worthwhile.

A good example is from the WHI study of diet and breast cancer that we have been discussing:

"There were departures from the design assumptions that likely reduced study power.

…

"If the WHI design assumptions are revised to take into account these departures [less dietary fat reduction], projections are that breast cancer incidence in the intervention group would be 8% to 9% lower than in the comparison group [and] the trial would be somewhat underpowered (projected power of approximately 60%) to detect a statistically significant difference, which is consistent with the observed results."

This illustrates the contorted sort of reasoning that speculation about low power requires.

What are they trying to say?

Their intended meaning seems to boil down to the following:

There might be a 9% reduction in risk. We could have missed it because power was only 60%.

This speculation is completely pointless, because the conclusion is better supported, and much clearer, from examination of the estimate and CI:

But HR = 0.91, so of course a 9% reduction is possible. It's what they actually saw!

Let's be clear on what happened here: world-class researchers reporting a very high-profile (and hugely expensive) study lapsed into convoluted and completely pointless reasoning. This is a dramatic illustration of how exclusive focus on whether $p<0.05$ leads to easily avoided problems. Had the authors followed BP2, no such speculation would have been required, and the implications of their results would have been much clearer.

BP2. Ensure that major conclusions reflect the estimates and the uncertainty around them.

The references given above on pages 4 and 5 are also relevant for this issue.

Problem 4. Exclusive reliance on intent-to-treat analysis This means analysis of all randomized subjects, regardless of how well they cooperated with treatment, possibly even including those who refused to actually undergo study treatment at all.

Intention to treat analysis is useful for preventing post-randomization self-selection from producing spurious positive findings, but it does not ensure the most accurate possible estimates for all purposes.

Consider a
'Negative' study of vitamin E in diabetics (*JAMA* 2005) *JAMA* **293**:1338-47, 2005
that claimed to have proven that vitamin E supplementation does not prevent cancer. (This was based on the p-value fallacy, by the way.) It used ITT:

"To reduce bias, we included continuing followup from those who declined active participation in the study extension and stopped taking the study medication."

But ITT produces underestimates of actual biological effects: it is biased toward no effect.

Thus, in addition to ignoring their estimates and CI's, they based a negative conclusion on an approach that is biased in that direction. This is very different from still having a positive finding despite some bias in the other direction, which is where ITT analysis works well.

This is an area where the WHI study did reasonably well. They used specialized methods to attempt to estimate what effect the fat lowering intervention would have had if it were followed as recommended. These methods try to avoid the self-selection bias that simple per-protocol analyses (or observational studies) would have, while also avoiding the biases of ITT analysis.

WHI: Estimate of effect if adherent to low-fat diet:
This estimate was:
Breast cancer HR 0.85 (0.71 – 1.02) This a bit lower than the 0.91 from the primary analysis, but it still just misses $p<0.05$.

They went further in trying to account for adherence to the intervention, but balked at giving any more details than the quote below:

Use of more stringent adherence definition "leads to even smaller HR estimates and to 95% CIs that exclude 1."

BP5. Learn as much as you can from your data.

Strictly limiting analyses to ITT only will sometimes not be enough to fulfill this goal.

Doing ITT analysis is usually important, so designing procedures to allow ITT is a good practice. In particular, it is good to continue to follow subjects who stop study medication. But ITT can be supplemented with additional analyses, notably analysis restricted to those who actually underwent the study treatments as planned, termed "per-protocol" analysis.

So a specific Best Practice is to

BP5a. Also consider per-protocol analyses, especially if:

- Interest in biological issues ITT is not designed to address biological effects and can be poor due to bias toward no effect
- Double-blinded treatment This reduces (but does not eliminate) the potential for self-selection biases that ITT protects against.

Having results from both ITT and per-protocol analyses can provide a fairer assessment of the uncertainty about a treatment's effect. This is especially important if negative conclusions from ITT analysis are not as well-supported or are contradicted by per-protocol analysis.

Also, when treatments are randomized and blinded, stratifying or controlling for the level of adherence or the time of dropout can produce an "in-between" estimate that may be sensible.

Another possibility is to

BP5b. Consider advanced methods to estimate causal effects.

There are new and complex "causal inference" methods that seek to avoid the biases of both simple ITT and per-protocol analysis. These are what WHI used. You are likely to need help from a statistician to carry these out.

Two closely related problems are:

Problem 5. Reliance on omnibus tests

Problem 6. Overuse of multiple comparisons adjustments

Omnibus tests (like ANOVA) are methods that

- check for any one or more of a large number of possible departures from a global null hypothesis (nothing is happening anywhere) They are
- inherently focused only on p-values (Problem 1) and they are
- diffuse, so weaker for specific issues

This makes them generally less useful than analyses focused on specific relationships whose magnitudes can be estimated as well as tested. In particular, when the p-value is large, the main use for omnibus tests is the misuse highlighted in Problem 1.

One reason that some people like omnibus tests is that they help guard against obtaining spurious positive results due to multiple comparisons. Because omnibus tests look broadly for any one of many possible departures from the null hypothesis, they are not good at finding any specific one. This makes them "conservative" for any specific question, which some people consider desirable or rigorous.

Multiple comparisons adjustments

- each result detracts from the other

Another way of guarding against chance false positive results is application of multiple comparisons adjustments. These are also inherently focused only on p-values, promoting use of the p-value fallacy. They also have the unfortunate property that the results of each analysis are automatically assumed to detract from all the others, with no consideration of how well the different results fit together conceptually or scientifically. Like omnibus tests, these are also very conservative, which some people like. But accuracy is a much more worthy goal than conservatism, and this is often better achieved by less formal (and more intelligent) ways of guarding against spurious findings.

The following page provides examples of how to justify not using multiple comparisons adjustments, both for papers and in grant proposals.

Example text for responding to a manuscript review or for inclusion in a paper:

> Although we examine many differences and issues, we report nominal p-values, without adjustment for multiple testing. Such adjustment would be focused on avoidance of one or more results with $p<0.05$ in the case where all differences are truly zero [Ref1-Ref3], which is an extremely unrealistic hypothesis about the state of nature in our situation. In addition, adjustment would require that each result detract from the others, but there are clear biological relationships among many of the issues that we examine, and these permit coherent sets of findings to reinforce each other rather than detract from one another. Thus, multiple comparison adjustment would do exactly the wrong thing in this case [Ref4]. We therefore rely on scientific judgment rather than formal adjustment methods to indicate where caution is warranted despite findings with $p<0.05$.
>
> Ref1. No adjustments are needed for multiple comparisons. K. J. Rothman. Epidemiology 1990: 1(1); 43-6.
> Ref2. Multiple comparisons and related issues in the interpretation of epidemiologic data. D. A. Savitz, A. F. Olshan. Am J Epidemiol 1995: 142(9); 904-8.
> Ref3. What's wrong with Bonferroni adjustments. T. V. Perneger. British Medical Journal 1998: 316(7139); 1236-8.
> Ref4. Bacchetti P. Peer review of statistics in medical research: the other problem. Br Med J, 324:1271-1273, 2002.

Example text for inclusion in a grant proposal:

> Although this Aim involves many different measures, we do not plan formal adjustments for multiple comparisons. This is because we expect many measures to show statistically significant differences, and that directions and magnitudes of differences (perhaps including some with $p>0.05$) will fit a biologically coherent pattern. In this case, each result will reinforce the other, rather than detracting from one another as required by formal multiple comparisons adjustments such as the Bonferroni method. Conversely, if only one or a very few measures reach statistical significance and their directions and/or magnitudes do not coherently fit with <<our substantive theory>>, then we will note that the result(s) with $p<0.05$ lack biological plausibility and could be due to chance despite meeting the conventional cutoff for statistical significance.

An investigator was very worried and puzzled.
Investigator's panicked inquiry:

He had done an
Animal experiment that included

- a condition that just confirms that the experiment was done correctly
- some places where different conditions should be similar
- some conditions that should differ

Saw expected results in pairwise comparisons, but "ANOVA says that there is nothing happening"

Because this had a specific focus on certain pairwise comparisons to address the scientific questions, he had done *t*-tests and estimated pairwise differences, obtaining positive results that he thought made sense. But he thought that he "had to" perform ANOVA, and this produced a p-value a bit larger than 0.05. So he thought that to be "rigorous" he would have to reach the opposite conclusion of what he found with the focused analyses. In fact, the focused results were what mattered; unfortunately, there is a risk that reviewers may think otherwise.

In fact, reviewers often state flatly that omnibus tests and multiple comparisons adjustments *must* be used when in fact those approaches would be very inappropriate.
Reviewer's comment on a study examining effects of 4 different administration routes This was a mainly descriptive study with many positive results, not a single positive result that was likely to be due to chance. But the reviewer stated:

"Repeated measures analysis of variance should be completed. Only if the time-by-treatment interaction is significant, should time-specific comparisons be made. Then multiple comparison procedures, such as Tukey's test, should be used rather than repeated t tests."

This would treat p>0.05 on the unfocused omnibus test of time-by-treatment interaction as a reliable indicator that no important differences are present—**Problem 1**.

The reviewer's comment, particularly the part highlighted, may sound rigorous, but it is only "rigorous" in the sense of being rigid or harsh, not in the sense of being exactly precise. It requires extreme conservatism—not accuracy—which could result in missing or understating important findings.

Another consultation concerned a study with a great deal of scientific structure that omnibus tests or multiple comparisons adjustments would not take into account. This was a

Study of biology of morphine addiction:

Very complex design involving:

- two different receptors
- antagonists
- different brain regions with and w/o certain receptor
- systemic vs local administration

Results of many pairwise comparisons fit a biologically coherent pattern.
Conditions that should have differed did, while comparisons that should have been similar were.

A reviewer of the manuscript ignored the consistency of the findings and wrote the following strident comment:

Reviewer: "The statistical analyses are naïve. The authors compute what appear to be literally dozens of t-tests without any adjustment to the alpha level --- indeed the probability of obtaining false positives grows with the number of such tests computed. The authors should have conducted ANOVAs followed by the appropriate post-hoc tests. Their decision to simple [*sic*] compute t-tests on all possible combinations of means is statistically unacceptable."

The highlighted statement is incorrect. The chance of obtaining *at least one* false positive increases *if* the null hypothesis holds for *all* comparisons. False positives in general do not become more likely, and the chance of getting many false positives that all fit together in a coherent biological theory is extremely small. This is a clear case where the results of multiple analyses all reinforce each other rather than detracting from each other as required by omnibus tests and multiple comparisons adjustments.

But the probability of obtaining multiple positive results exactly where expected and negative results exactly where expected does not grow; it becomes vanishingly small.

Striving for the following best practices will often lead to much better analyses and interpretations than use of omnibus tests and multiple comparisons adjustments.

BP6. Base interpretations on a synthesis of statistical results with scientific considerations.

In clinical research, there is usually outside knowledge that can be used to help with the choice of analyses and their interpretation. Recognizing and explaining whether and how results of different analyses fit together is crucial for obtaining the best understanding of what can be learned from the study. This will usually require consideration of the directions and magnitudes of estimated effects, along with the uncertainty shown by the CI's, rather than consideration of p-values alone.

BP6a. Rely on scientific considerations to guard against overinterpretation of isolated findings with $p<0.05$. (This is usually preferable to formal multiple comparisons adjustment.)

In particular, it is important to realize when one or a few findings reach $p<0.05$ but the ensemble of results does not have a compelling explanation. If the results with $p<0.05$ are not especially more plausible than other quantities estimated, and the directions and magnitudes of these and other results do not show patterns that reinforce the findings, then it is reasonable to regard those findings as suggestive (or even potentially spurious) rather than conclusive, despite their small p-values. Given that our publishing environment has substantial disincentives for such cautious interpretation of findings with $p<0.05$, this requires strong dedication to fair interpretation.

BP6b. Acknowledge the desirability of independent replication, particularly for unexpected findings.

Replication plays a crucial role in solidifying scientific knowledge, but the tendency to focus only on supposedly conclusive "findings" can sometimes cause this to be overlooked. For example, a recent editorial, by a prominent statistician, addressed the topic of subgroup analyses, which is a version of the multiple comparisons issue (Wittes J. On Looking at Subgroups. *Circulation*, **119**:912-915, 2009). It failed to mention any role for replication.

BP7. Choose accuracy over conservatism whenever possible.

Many consider conservatism to be very desirable and rigorous, but this certainly is not so when accuracy is a viable alternative. Conservatism is a type of bias, and bias is bad. Sometimes it is better to know the direction of the bias rather than to be uncertain. Intent-to-treat analysis, omnibus tests, and multiple comparisons adjustments introduce bias with a known direction, but it is still bias. You will often be able to do better by thinking carefully about all your results.

BP1 and BP2, obtaining estimates and CI's and taking them into account when interpreting results, will often be helpful for achieving accurate interpretation. They will usually steer you away from overly conservative, automatic methods based only on p-values.

Problem 7. Entangled outcomes and predictors

Predictors and outcomes that have parts of their definitions in common can cause severe problems.

Body mass index as a predictor of central fat

In the case of fat abnormalities in HIV infected persons, early studies controlled for BMI when assessing whether peripheral fat loss was associated with central fat gain. Because fat amounts contribute directly to BMI, this caused a spurious association.

Many people have low peripheral and low central fat

A few (with HIV and not) have low peripheral fat and high central fat

Those with low peripheral fat were less likely to have high central fat. But the opposite picture emerges if you control for BMI.

Low peripheral fat + low central fat → low BMI

So low central fat "explained" by low BMI in these cases

Association of peripheral fat and central fat therefore determined by rare cases of low peripheral fat and high central fat, causing a spurious association

Those with low values of both peripheral and central fat had their low central fat "explained" by a low BMI, thereby leaving the estimated association to be determined mainly by the rare cases with low peripheral fat and high central fat. This field of inquiry was seriously misguided for many years because of this seemingly obvious problem.

Total time on treatment as a (fixed) predictor of survival time

Can only be treated if alive

(survival time = treated time + untreated time) This equation shows the entanglement, with the predictor "treated time" contributing directly to survival time.

Died after 2 days → max of 2 days treatment

Treated for 5 years → min of 5 years survival

Meaningless association

Those who die early will necessarily have their total time on treatment limited by how long they lived. Someone who died after 2 days will at most have 2 days of treatment. Conversely, someone with 5 years of treatment must have lived at least 5 years. Proportion of time on treatment, although less obviously biased, also may depend on survival time and be potentially biased. Another example might be number of rejection episodes as a predictor of post-transplant survival. Someone who died on day 1 would have at most 1 rejection episode, while someone with 5 episodes would have to have lived quite a bit longer.

To avoid these problems

Either

1) ensure that outcome is not part of the definition of a predictor, and vice versa, or

2) be very careful and clear with interpretation

In general, it is simplest and safest to make sure that outcome and predictors are fully distinct.

For time-to-event analyses

Fixed predictors should be defined without using any information from after the start of followup

Use time-dependent covariates, defined only using measurements up to present

In general, any fixed (non-time-dependent) covariates should be known from information available at time zero. And for time-dependent covariates, any use of “look-ahead” information is likely to cause trouble.

Some technical problems and ways to avoid them

Techincal problems

Unchecked assumptions

The lectures and labs have emphasized this. Be sure to describe what you did to check assumptions and what you found.

Ignoring dependence and clustering

An extreme case of this is performing unpaired analyses on paired data. The upcoming lectures will deal with situations that have clustering or repeated measures.

Unclear details for time-to-event: operational definitions, early loss, event ascertainment

Provide operational definitions of starting time, occurrence of event, and censoring time. Any analysis of time to an event needs to be clear about the time from when to when.

Summarize followup among those who were censored, because followup is complete for anyone who had the event; it does not matter whether the event occurred at 2 days or 5 years—either way, we know all we need about that person's outcome. The amount of followup matters for those who did not have the event, especially the minimum followup and/or the number of subjects with shorter than desired followup. Mixing the early events into summaries of followup times obscures this information.

Summarize early loss to followup and the reasons for it. Censoring due to loss to followup is more likely to violate the assumption of non-informative censoring, so this is a particular concern that should be addressed separately from observations censored just due to the planned end of the study or observation period.

Describe how events were ascertained This is important for establishing the completeness of ascertainment, and sometimes for explaining clumps of events (e.g., if many were found at a scheduled 6 month visit).

Missing data

See annotated slide on page 32, below, for discussion of this issue.

Poor summaries (e.g., mean±SD for skewed data)

For skewed data, medians are usually better summaries. The CI for the median can be added to show how precisely it is estimated, or the range and/or quartiles (IQR) can be added to show the variation in the population.

There is sometimes confusion about whether to show SD's or SE's. Show SE's (or CI's, usually better) to indicate the precision of estimates. Show SD's to indicate the variability in the population.

Showing inadequate or excessive precision

In general, too little precision may leave the reader wondering about the exact magnitude. For example, p=0.01 could mean anything from 0.005 to 0.015, which is a pretty wide range (3-fold). Extra precision is not directly harmful, but gives a spurious impression of how precise the results are. It also can look naïve, giving an astute reader or reviewer the impression that you do not know what is important and what is not.

For odds ratios, relative risks, and hazard ratios, I suggest one of the following: 1) give the value to two decimals if <2.0, one if >2.0, or 2) give all values to two decimals. The second has the possible advantage of always the same number of decimals, but this can be excessive. OR's like 24.56 look a bit odd.

Give p-values to two significant digits (leading 0's don't count), to a maximum of three or four. Sometime people give a maximum of three digits, and this is what some Stata procedures provide. But giving up to four is sometimes desirable and is usually also fine.

Do not use "P<" for values of 0.001 or more; use "P=". For example, don't say p<0.01 when you could say p=0.0058.

Never use "p=NS" or "p>0.1" This gives needlessly vague information and encourages Problem 1.

Do not show χ^2 or other statistics that provide the same information as p-values (but are less interpretable). These add no information and clutter presentation of results. They may seem to add some technical cachet, but leaving out unimportant details actually conveys a better impression of technical savvy (to me, at least).

Poorly scaled predictors

Common examples of this problem include using age in years or raw CD4+ cell count. In regression models, the coefficients for numeric predictors are the estimated effects of a 1-unit increase in the predictor. So if the age variable is in years, the estimated effect is for a 1-year increase in age, which is often too small to be readily interpretable. When a 1-unit increase is a very small amount, estimated coefficients will necessarily be very small and results will be hard to interpret.

For example, an OR of 1.0051 per 1 cell/mm^3 increase in CD4 count is very hard to interpret. It is also hard to rescale this by eye, because the OR for a 100 cell increase is $(1.0051)^{100}$, which most people cannot calculate in their heads (it's 1.66).

To avoid this problem, rescale numeric predictors before running regression models (or use lincom afterwards). For example, make a new variable cd4per100=cd4/100.

Terms likely to be misread ("significant")

Don't needlessly give readers and reviewers the opportunity to misunderstand what you mean.

Avoid use of "significant" alone. Use "statistically significant" if meaning p<0.05; use "important", "substantial", or "clinically significant" if that is the intended meaning. As noted under Problem #1 (page 5, above), some journals reserve "significant" alone to mean "statistically significant". If they do not allow the full term, just avoid using it at all (this may be a good strategy anyway).

Use "Relative Hazard" or "Hazard Ratio" for proportional hazards model results, not "Relative Risk"; this is sometimes used for analysis of a binary outcome.

Use "Mann-Whitney" instead of "Wilcoxon" or "Wilcoxon rank-sum" to avoid confusion with "Wilcoxon signed-rank". Both "Mann-Whitney" and "Wilcoxon rank-sum" are used, due to the near-simultaneous, independent development of the method.

Missing Data

Can cause bias, depending on why missing Some causes of missing data can produce severe bias in the data that remain. For example, those with the riskiest behavior may be more inclined to skip questions about risk factors. The more that is missing, the more likely that there is a problem with what remains. Think about why participants may have declined to provide the information.

Prevention is best; otherwise

- Clearly disclose how much
- Give reasons when known
- **Assess differences in non-responders** You can compare those who responded to those who didn't on demographics and other variables that are more complete. This can provide clues about why some people didn't respond and what impact that may have on your results.
- **Perform sensitivity analyses** For example, when comparing two treatments, you can make a pessimistic assumption about those with missing outcomes in one arm and an optimistic assumption for the other arm. If the conclusion remains qualititavely the same, then you can have confidence that it was not caused by bias due to missing data. In some cases, however, conclusions may reverse, leaving doubt.
- **Consider advanced methods** For example, Stata can facilitate a method called *multiple imputation* (see the *mi* command and http://www.ats.ucla.edu/stat/stata/Library/ice.htm). You will likely need help from a statistician to use such approaches.

When comparing univariate and multivariate results, use same set of observations.

For multivariate modeling, the default in Stata and other programs is to delete all observations that have missing values for any predictor or for the outcome. (For stepwise selection, observations missing any *candidate* predictor are deleted, even if the candidate is never used.) This means that smaller models may have more observations than larger ones. Keep this in mind if you need to compare the results of such models, such as when you are examining how an effect of interest changes when controlled for other predictors. In these cases, you should fit the models to the same set of observations, even though the smaller models could utilize more.

Homework

Examine the two assigned papers

Look for:

- use of best practices, other strengths
- problems
- missed opportunities for using best practices

Think about what would have been better and the practical or scientific consequences

We will discuss these on Thursday

Written Homework
Due before class on Thursday 4/19

Choose one paper and

> Describe occurrence of one of the problems. Quote text or pinpoint specific results or passages to clearly identify what you discuss. What should have been done to avoid the problem?

Heisler M, Faul JD, Hayward RA, Langa KM, Blaum C, Weir D. Mechanisms for Racial and Ethnic Disparities in Glycemic Control in Middle-aged and Older Americans in the Health and Retirement Study. *Arch Intern Med* 2007; **167**: 1853-1860.

Homsy J, Bunnell R, Moore D, King R, Malamba S, et al. Reproductive Intentions and Outcomes among Women on Antiretroviral Therapy in Rural Uganda: A Prospective Cohort Study. *PLoS ONE* 2009; **4**(1): e4149.

Here are the Problems and Best Practices gathered together for easy reference.

Summary of Problems

Problem 1. P-values for establishing negative conclusions
Problem 2. Misleading and vague phrasing
Problem 3. Speculation about low power
Problem 4. Exclusive reliance on intent-to-treat analysis
Problem 5. Reliance on omnibus tests
Problem 6. Overuse of multiple comparisons adjustments
Problem 7. Entangled outcomes and predictors

Summary of Biostatistical Best Practices

BP1. Provide estimates—with confidence intervals—that directly address the issues of interest.
BP2. Ensure that major conclusions reflect the estimates and the uncertainty around them.
BP2a. Never interpret large p-values as establishing negative conclusions.
BP3. Discuss the implications of your findings for what may be true in general. Do not focus on "statistical significance" as if it were an end in itself.
BP4. State what you did find or learn, not what you didn't.
BP5. Learn as much as you can from your data.
BP5a. Also consider per-protocol analyses, especially if:

- Interest in biological issues
- Double-blinded treatment

BP5b. Consider advanced methods to estimate causal effects.
BP6a. Rely on scientific considerations to guard against overinterpretation of findings with $p<0.05$.
BP6b. Acknowledge the desirability of independent replication, particularly for unexpected findings.
BP7. Choose accuracy over conservatism whenever possible.

Specific exercise for written projects: Perform the checking procedure described on Tuesday on your own paper.

Perform searches for words "no" and "not" Whole word searches on these two terms should find most negative interpretations of statistical analyses.

Check each sentence found

- Is there an estimate and CI supporting this?
- What if the point estimate were exactly right? Would the conclusion still make sense?
- What if the upper confidence bound were true? Does the conclusion allow for this possibility?
- What if the lower confidence bound were true? Does the conclusion allow for this possibility?

Additional searches: "failed", "lack", "absence", "disappeared", "only", "rather", "neither", "none" Negative interpretations sometimes use these words, so you can also check them to make sure you didn't miss anything.

Also for your written projects

Try to avoid the other problems and follow the best practices

(Or be clear on why your case is an exception)

The best practices may seem fairly obvious and uncontroversial, even just common sense. Unfortunately, many people have been taught that statistical reasoning contradicts and overrules common sense. In particular, some reviewers or editors may tell you that you cannot pay any attention to estimates when $p>0.05$. But a complete assessment of a study's implications will usually require consideration of estimates, both direction and magnitude, and this will usually be acceptable as long as you do not downplay uncertainty or exaggerate the conclusiveness of your results. I encourage you to follow these practices in order to obtain the best assessment and presentation of what can be learned from your data.

Take advantage of the faculty help that is available

Please remember that you have faculty help available and we are eager to help however we can.