Common Biostatistical Problems
And the Best Practices That Prevent Them

Peter Bacchetti
Biostatistics 209
May 16, 2006

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 interpretation and presentation of findings

You may have seen much 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, following these guidelines is surprisingly difficult, even when you understand the principles behind them.

Your class projects are 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

Please let me know about additions or disagreements

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

Problem 1. P-values for establishing negative results

This is a huge problem, not just the first among equals. This is very common in medical research and leads to terrible misinterpretations.

The P-value Fallacy (Fantasy): The term "P-value fallacy" has been used to describe more subtle misunderstandings, so the term "fantasy" may be a better choice.

Almost no one would really defend the first two statements below. They 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.

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.

But the statement below may seem a bit more defensible, because it resembles what people are 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.

If the result is not statistically significant, you "have to" conclude that there is no difference.
And you certainly can't state that there is any suggestion of an effect.

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".

Here is an example. This is fairly relevant because it is closely based on a student project for this class from a previous year (but with some rounding, etc., to make it anonymous).

A randomized clinical trial of 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 p=0.13. This type of interpretation is alarmingly common, but the difference is not zero ("no difference"); it is 8%.

Sometimes you instead see reports like those below:

"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.

Finding egregious examples of this fallacy in prominent places is all too easy. For example, I looked at two recent issues of *NEJM*, and this was the first 'negative' study I found:

April 27, 2006 *NEJM*, **354**: 1796-1806

"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 what may be the most important outcome has been interpreted as definitive evidence against effectiveness. In fact, if we new that this observed estimate was correct, then vitamin supplementation, or at least further study, would probably be worthwhile.

A less blatant but even higher-profile example is provided by the recent report on the
Women's Health Initiative study on fat consumption and breast cancer

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

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—or were forced—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 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)

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.

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

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.

There is 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 a recent example of a strong negative result that is well supported

May 4, 2006 *NEJM*, **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:

"Vitamin C and E supplementation appeared to reduce 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 great to permit a strong conclusion that any protective effect exists. Unfortunately, this more accurate interpretation would probably have greatly reduced the paper's chance of acceptance at *NEJM*. We all know that there is a lot of pressure to make results seem as interesting as possible, but this should only go so far. Using the p-value fallacy to make a study seem definitive in one direction instead of 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).

Following directly from BP2 is the following more specific guideline:

BP2a. Never interpret large p-values as establishing negative results.

What if supported by a power calculation?

Still no good!

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

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

Some references are:

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.

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 exist in the real world. Interpretation should focus clearly on what evidence the study provides about what may be 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 point is often forgotten and we slip into thinking that statistical significance itself is what really matters. Most people understand that statistical significance implies strong evidence for a real 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 BP3 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. 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 difference may reach $p<0.05$, not that it 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. Last year, for example, most written projects for this class did contain lapses.

Easy to slip into relying on "p>" reasoning

- 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.

Be vigilant

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

Your projects for this class are a good opportunity to practice avoiding the p-value fallacy.

Exercise for written projects:

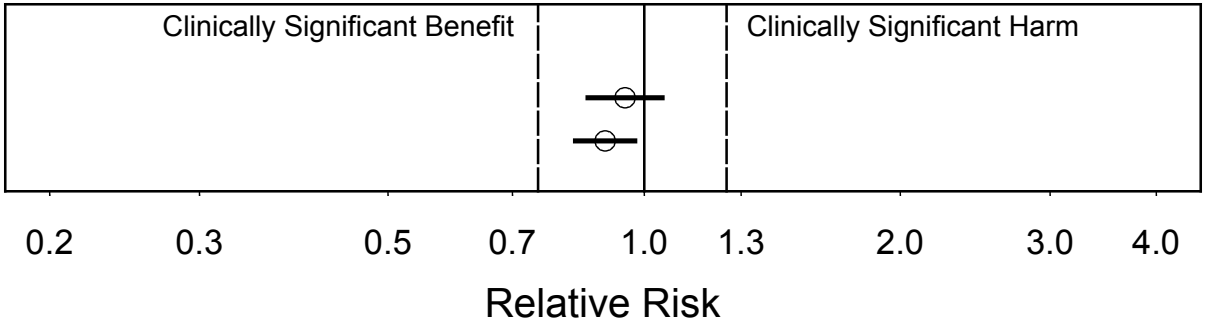
Perform searches for words "no" and "not" Whole word searches on these two terms should find any negative interpretations of statistical analyses.

Check each sentence found

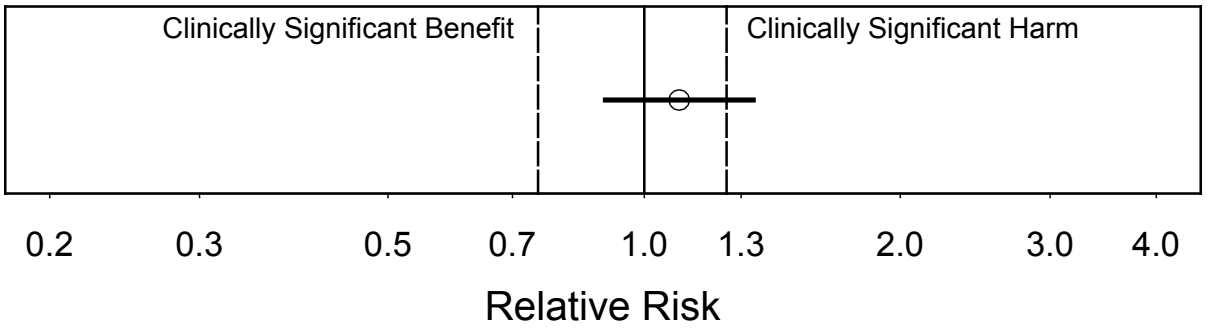
- Is there an estimate and CI supporting this?
- What if the point estimate were exactly right?
- What if the upper confidence bound were true?
- What if the lower confidence bound were true?

Additional searches: "failed", "lack", "absence", "disappeared" Negative interpretations sometimes use these words, so you can also check them to make sure you didn't miss anything.

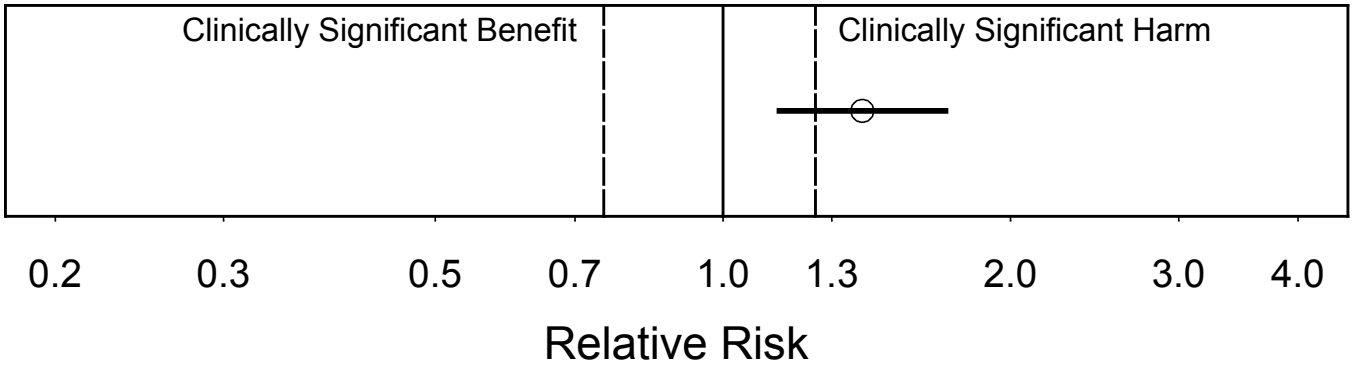
The following figures show some concrete examples of how to interpret estimates and CI's. These assume a somewhat idealized situation where we have exact limits on what is clinically important, but they illustrate the main ideas. Often it will be more practical to first calculate the estimates and CI's 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.



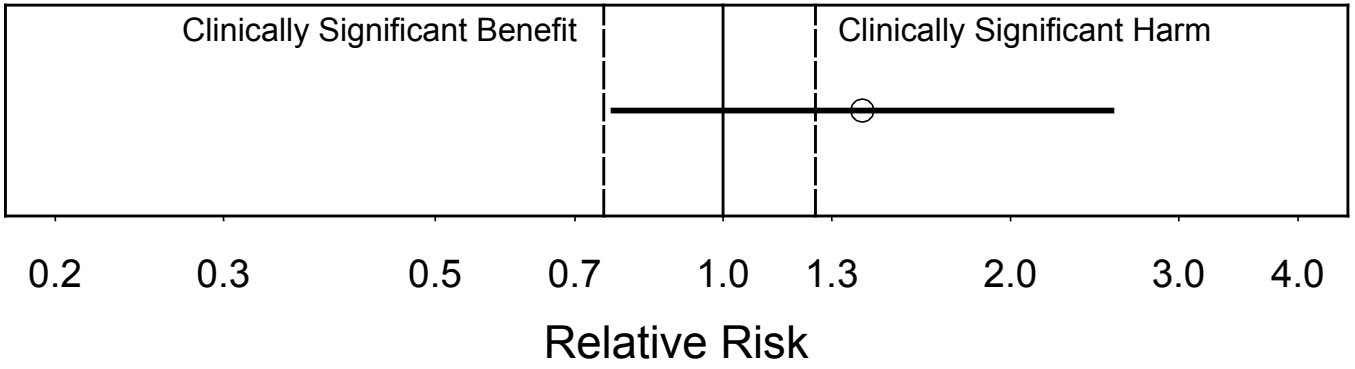
Strong evidence against any substantial harm or benefit Because we have strong evidence against any values outside the CI, both these cases argue strongly that any effect is clinically unimportant. Note that this is true even though one is statistically significant.



May be harmful The estimate is in the harmful range, with the CI including some values that would be clinically important harm.
Strong evidence against any substantial benefit The CI does not include any values that would be an important benefit.

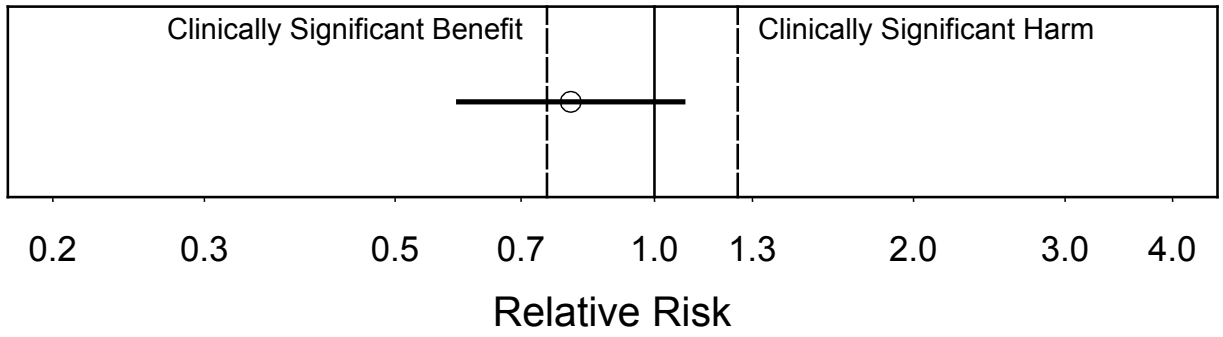


Likely to be harmful The estimate and most of the CI are in the substantial harm range, and the CI does not include no effect.

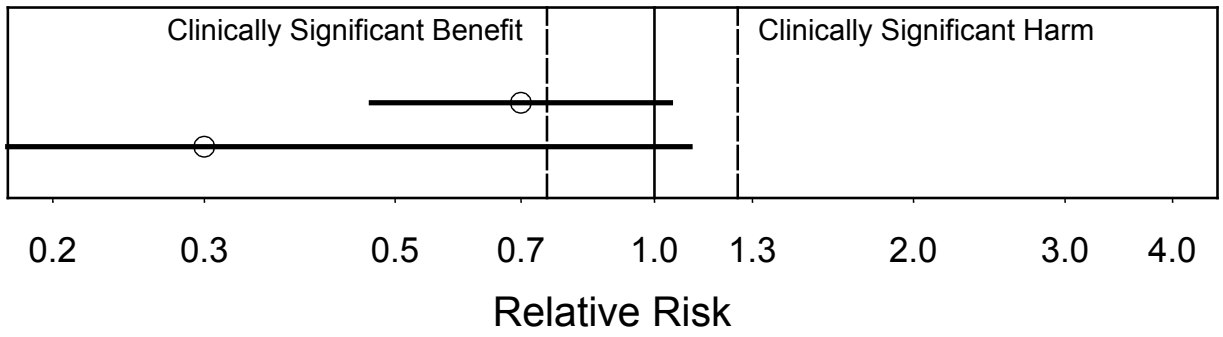


Results suggest substantial harm, but CI is too wide to permit a strong conclusion The estimate is in the substantial harm range, but the CI includes less important harm, no effect, and clinically insignificant benefit.

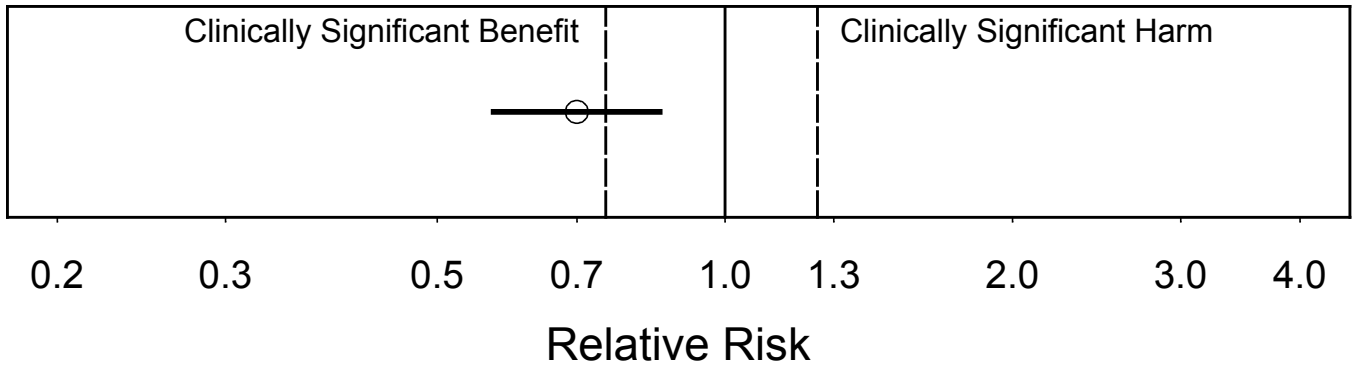
Strong evidence against substantial benefit The CI does not include any values that would be an important benefit



Suggestion of benefit Estimate and most of CI are in the benefit range
Clinically meaningful benefit possible but not likely Some of the CI is in this range, but not the estimate
Strong evidence against substantial harm No values in the CI would constitute important harm
May be no effect (not statistically significant) The CI includes no effect

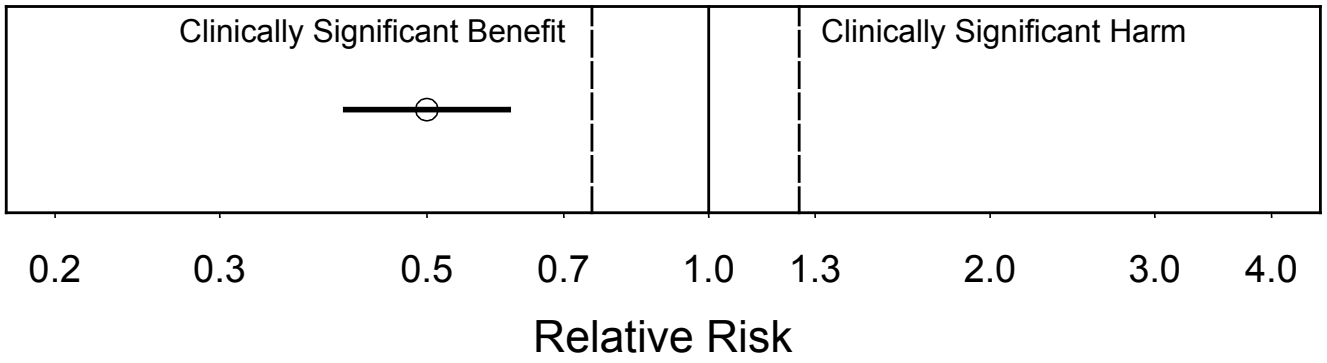


Suggestion of substantial benefit The estimate would be an important benefit if true
May be no effect (not statistically significant) The CI includes no effect
Which of the two results would be more exciting? I think the lower one is, even though it has wider uncertainty, because the estimate is better.

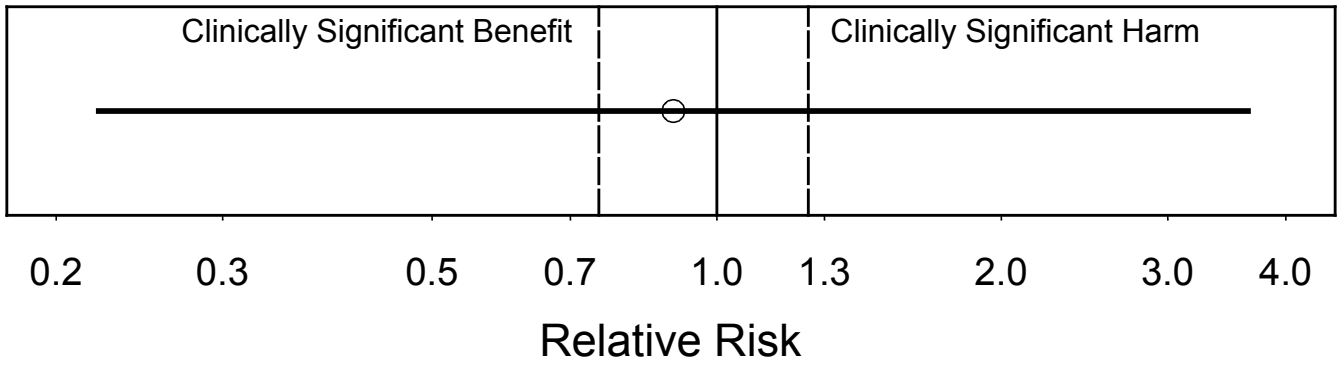


Strong evidence of benefit (statistically significant)

Substantial benefit appears likely, but CI too wide to rule out clinically insignificant benefit The CI includes some benefits that would be too small to be clinically important. Studies almost never include this observation if they reach statistical significance.



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 p. 64 of Vittinghoff, et al. textbook) This gives some more detailed verbal descriptions of some types of results.

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

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

Checks for possible assumption violations when little suspicion This is such a situation

Just need to be able to state that you checked and nothing jumped out; don't need to prove that nothing was present 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." 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. The second forms are based on the p-value fallacy.

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 …

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 invite problem #1, above. When there is strong evidence against an effect, they are too weak.

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

-- Prime Minister John Major, 1995

Of course, it turned out that BSE *was* transmitted to humans, and over 150 people have 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.

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 are the phrases that first pop into my mind when I think about how to state a finding.

FRAM, nationwide study of fat abnormalities in 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 (VAT) increased as peripheral fat decreased and the two were causally linked.

Despite the strong results, some phrasing in an early draft was:
Original: "our results do not support the existence of a single syndrome with reciprocal findings."

We did catch this and revise it 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

Among many results supporting this conclusion were the following, showing that peripheral fat loss did not have any substantial association with central fat gain (note upper confidence bound) and that it did have a strong association with central fat loss.
Central hypertrophy association with peripheral lipoatrophy, OR: 0.71, CI: 0.47 to 1.06, P = 0.10.

Central lipoatrophy association with peripheral lipoatrophy, OR: 18.9, CI: 5.7 to 63, P < 0.0001.

Another example:
Safety of cannabinoids in persons with treated HIV

Original: "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

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.

This problem seems very common. The last manuscript I edited, just last week, also used phrasing that was weaker than it needed to be.

Last week:

(This is not yet published, so I haven't given any substantive details.)

Original: "our findings do not support previous reports of an association …"

Suggested: "our findings contradict previous reports of an association …"

Observed association is in the opposite direction of previous reports and is statistically significant.

This study avoided some possible biases in the previous studies, and its CI rules out the previously claimed association. So it does "contradict" the previous reports.

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 under 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 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 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?

I believe that their intended meaning boils 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 a cursory examination of the estimate and CI:

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

Had the authors followed BP2, no such speculation would have been required, so it deserves an encore:

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

Some relevant references are given above on page 8. Note also that the CONSORT document concerning guidelines for reporting clinical trials states that "There is little merit in calculating statistical power once the results of the trial are known" (*Ann Int Med* 2001, 134:663-694).

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.

An example of Problem 1 from around this time last year was a
'Negative' study of vitamin E in diabetics (*JAMA* 2005) *JAMA* **293**:1338-47, 2005
that claimed (based on the p-value fallacy) to have proven that vitamin E supplementation does not prevent cancer. 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 the effect that the fat lowering intervention would have if followed as recommended (and intended by the study). 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:

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, such as continuing to follow subject who stop study medication, is a good practice. But ITT can be supplemented with additional analyses, notably analysis restricted to those who actually underwent the study treatments, termed "per-protocol" analysis.

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

- Interest in biological issues ITT is not designed to address these 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 less well supported by per-protocol analysis.

BP5b. Consider advanced methods to estimate causal effects.

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

Problem 5. Reliance on omnibus tests
Problem 6. Overuse of multiple comparisons adjustments

These two problems are closely related

Omnibus tests (like ANOVA)

- inherently focused only on p-values (Problem 1)
- diffuse, so weaker for specific issues

Omnibus tests—those that test for any one of many different possible departures from a global null hypothesis—are inherently focused only on p-values and do not give any focused analysis of 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.

A few months ago, I got a call from an investigator who was very worried and puzzled.

Investigator's panicked inquiry:

Animal experiment that included

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

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. I persuaded him that the focused results were what mattered, but warned him that reviewers may think otherwise.

In fact, I very often see comments from reviewers stating 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 recent study examining effects of 4 different administration routes This was a mainly descriptive study with many positive results, not a single one that was likely to be due to chance.

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

Study of biology of morphine addiction:

Very complex design involving:

- two different receptors
- antagonists
- different 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 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.

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 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.

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.)

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 will often be helpful for achieving accurate interpretation, so it deserves an encore.

BP1. Provide estimates—with confidence intervals—that directly address the issues of interest.

Obtaining estimates will steer you away from automatic methods based only on p-values.

This very fundamental problem is surprisingly easy to slip into.

Problem 7. Reversed predictors and outcomes

Example: Survival to discharge following MI This is just a hypothetical example

Variable	Survived (N=200)	Died (N=50)	P-value
Age (mean±SD)	62±7	71±12	<0.0001*
Sex: Male	150 (68%)	45 (87%)	0.0096†
Female	70 (32%)	7 (13%)	
…	…	…	…

This shows a table that would be more appropriate if survived versus died were the predictor and the row variables were outcomes. For example, the percentages do not directly show how much higher the risk of death is for males compared to females.

*Unpaired t-test with unequal variances
†Fisher's exact test

The right way around:

Univariate logistic regression models

Predictor	# Died/N (%)	OR (95% CI)	P-Value
Age (per decade)		2.2 (1.53-3.2)	<0.0001
Sex: Female	7/77 (9)	--reference--	
Male	45/195 (23)	3.0 (1.25-8.3)	0.0096
…	…	…	…

This shows how to analyze the same data the other way around, with the outcome and predictors in the right roles. We can see directly the estimated impact of age and sex on risk of death, including the more useful percentages.

Another encore for BP1.

BP1. Provide estimates—with confidence intervals—that directly address the issues of interest.

In particular, you can avoid this problem by considering whether reversing your analyses would more directly show what you need to address the key issues. In tables showing descriptive summaries, this can also help determine which percentage (row or column) is most helpful.

Problem 8. Outcome transformations that hinder interpretation

Transforming outcome variables can be very useful for reducing outliers and improving the validity of standard statistical methods. But care is needed to ensure that results are still interpretable.

Example: Rank transformation. "Women averaged 10 ranks higher than men on depression scores."

Replacing numeric values of an outcome variable with their ranks may eliminate outliers and make the data look more normal, but any directly interpretable quantitative information is lost. This will usually limit the information obtained to p-values, with no meaningful estimates or CI's, contrary to BP1-2. In contrast, logarithmic transformation often helps with skewness and preserves the potential for quantitative interpretation. When the outcome has been modeled by multiple regression after logarithmic transformation, exp(_) can be interpreted as a fold-effect, and 100*(exp(_)-1) is a percentage effect (both multiplicative).

Example: Analysis of cost data on logarithmic scale. "The estimated effect of a major complication was a 2.4-fold increase in cost."

In some cases, logarithmic transformation is not appropriate. Cost data are often skewed and much better behaved after logarithmic transformation, but this makes observations with large costs less influential in subsequent analyses. When the focus is on resource utilization, a patient with $100,000 in costs really does have a much larger impact on the bottom line than one with $5000 in costs. This much greater impact makes standard statistical methods invalid, but deflating this impact by taking logs is not appropriate.

BP1a: Analyze outcomes on a meaningful, interpretable scale whenever possible. (Consider use of bootstrapping methods to obtain valid analyses when assumptions for standard methods are violated.)

This is an adjunct to BP1 because it is key for directly addressing the issues of interest.

Ideally, modeling the right outcome, measured in the right way, should be a top priority even when it causes some statistical difficulties. Computationally intensive bootstrapping methods can often produce valid confidence intervals when standard methods cannot. Stata has a tool for facilitating this approach. Consultation with a statistical expert will usually be warranted when such difficulties arise.

Technical Issues

Unchecked assumptions

- Normality (of residuals), outliers

It can happen that an outcome variable appears non-normal, but predictors in a multiple regression model can explain enough so that the remaining unexplained residuals appear normal. The Stata sktest command can check normality. T-tests technically are based on a normality assumption, but outliers are what really mess them up. Plots are often helpful for showing outliers. Deleting outliers, or changing them to smaller values, can be very dangerous. This changes the data and may introduce bias, so other strategies are usually preferable (transformation or bootstrapping). It may be OK to report results based on deleting or changing outliers as confirmatory analyses.

- Linearity for numeric predictors

The linearity assumption can be checked by adding the square of the predictor to the model, or by breaking the predictor into categories. Scatterplots and smoothing methods are also useful.

- Proportional hazards assumption

The Stata stphtest and stphplot commands can be used to check this assumption. Covariates that appear to violate proportionality can be controlled by stratification, or the non-proportionality can be modeled using time-dependent covariates defined as the product of the covariate with time itself (or log of time itself).

- Interactions

Check these by adding interaction terms to the model. Usually, only suspected or plausible two-way interactions are examined, because the number of interactions is so large (particularly if 3-way interactions are considered) and they are hard to estimate accurately.

You've learned in this class about this problem and methods for avoiding it.

Ignoring dependence in the data

- Unpaired summaries and tests for paired data

An elementary mistake, but not unheard of. I once had clients who were very skeptical of my insistence that unpaired analyses were inappropriate, because that was what had always been done for similar experiments in their area of investigation.

- Ignoring clustering (by provider, hospital, etc)

Outcomes among different patients in a study may not always be completely independent as assumed by simple methods. Different patients of the same provider may fare more alike than patients from two different providers. In fact, this will usually be the case, because providers do not give completely standardized care. Likewise, there are often hospital-to-hospital differences, or many other possible sources of dependence or clustering.

- Repeated measures

Different measures from the same person are usually more alike than measures from two different people. And measures closer together in time are often more alike than measures farther apart in time.

Poor description of survival analyses

Provide:

- Operational definitions of starting time, occurrence of event, and censoring time

This is neglected with surprising frequency. Any analysis of time to an event needs to be clear about the time from when to when.

- 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).

- Summaries of followup among those censored

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.

- Summaries of early loss to followup and reasons

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.

Mean ± SD when data are non-normal

Readers will tend to interpret these summaries as if the data were normal, so using these summaries in other situations can produce confusion.

Use median and range (and quartiles)

These are often acceptable for summarizing non-normal values in a study population

Or geometric mean with CI

This is sometimes a useful summary of non-normal data; it is based on log-transformed values (log transform, take mean, then take antilog to get the geometric mean on the original scale)

SD vs SE

Although there is often confusion about which is appropriate to present, these address different issues:

SD to show variability in a population
SE to show uncertainty around an estimate

Pick the one that shows what matters to the point you want to make.

Too little or too much precision

OR=0.3, OR=2.537
P=0.01, P=0.4275

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. 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.

Give OR's to two decimals if <2.0, one if >2.0

This may err on the side of giving too much precision in that smaller OR's (say, down to 1.5) are often given to only one decimal, but I think this is a reasonable proposal. The same would apply for relative risks and hazard ratios.

Give p-values to two significant digits (leading 0's don't count), to a maximum of four.

Sometime people instead give a maximum of three digits, and this is what some Stata procedures provide. This is usually also fine.

Do not use P< for values of 0.0001 or more; use P=.

That is, don't say p<0.01 when you could say p=0.0058.

P=0.13, P=0.013, P=0.0013, P=0.0001, P<0.0001

This also may err on the side of sometimes giving a little more precision than is needed, but it is not too excessive. P-values are often limited to three decimals, with p<0.001 being the smallest reported. This may sometimes be unavoidable when software only produces 3 decimals.

Never use "p=NS" or "p>0.1"

This gives needlessly vague information and encourages Problem 1.

Do not show $_^2$ 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 numeric predictors

Age in years

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.

These give estimated OR's (or RR's or HR's) very close to 1.0.

OR 1.0051 for each 1 cell/mm^3 increase

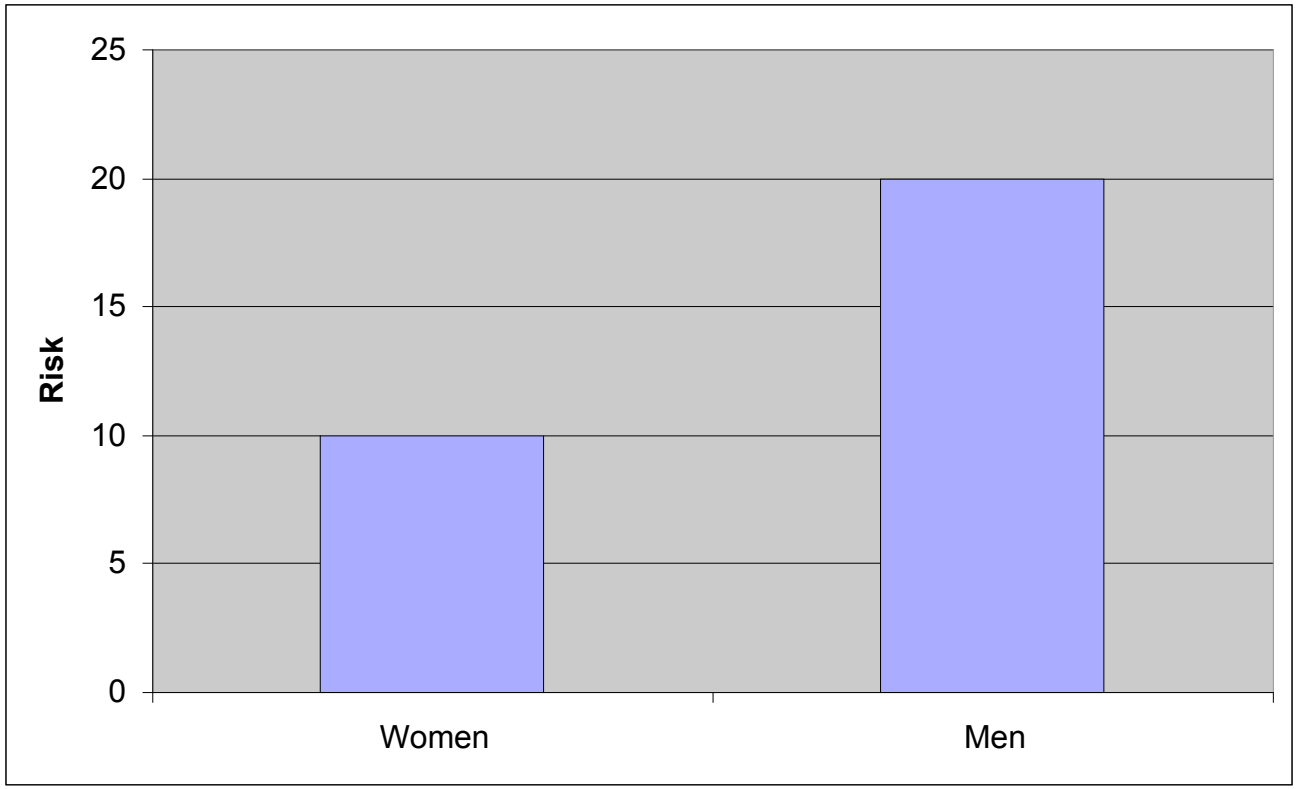
An OR of 1.0051 per 1 cell/mm^3 increase in CD4 count is very hard to interpret. It is also hard to rescale this by eye, because the OR for a 100 cell increase is $(1.0051)^{100}$, which most people cannot calculate in their heads.

Rescale numeric predictors to make results interpretable

Age in decades (effect per 10 year increase in age) Make a new variable age10=age/10 and use that as a predictor

CD4/100 (effect per 100 cell increase) Make a new variable cd4per100=cd4/100

Figures with low information content



This is an extreme example: the graph only shows two numbers. In some fields, low-info figures may be necessary for clarity or visual impact, but including more information is still usually preferable. For example, add confidence interval bars and p-values, or find a way to show a cross-classification all on one graph. The best use of figures is when they can clearly show information that is impractical to give in text or tables.

Terms likely to be misread Don't needlessly give readers and reviewers the opportunity to misunderstand what you mean

Use "Mann-Whitney" instead of "Wilcoxon" or "Wilcoxon rank-sum"

- Could be confused with "Wilcoxon signed-rank"

Both "Mann-Whitney" and "Wilcoxon rank-sum" are used, due to the near-simultaneous, independent development of the method.

Use "Relative Hazard" or "Hazard Ratio" instead of "Relative Risk" for proportional hazards model results

- Could be confused with analysis of a binary outcome

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, 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).

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.

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.

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

BP7. Choose accuracy over conservatism whenever possible.

These 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.

For your written projects

Try to avoid these problems and follow these guidelines

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

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.