This document consists of some of the annotated lecture notes from a 2008 lecture by Peter Bacchetti on "Common Biostatistical Problems and the Best Practices that Avoid Them." Smaller text is commentary on the actual slide text, which is shown in larger type. The lecture will be given again in April 2009; send any suggestions to peter@biostat.ucsf.edu

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.

Almost no one 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.

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 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.

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".

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.

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 have to be based on assumptions that are hard to know in advance.

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 calculating the statistical power once the results of the trial are known".

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T, for the CONSORT Group. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;**134**:663–94. Page 670

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.

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. This is fairly relevant because it is closely based on a student project for this class from a previous year (but with some rounding and so on 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 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”; it is 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 Medscape article: Statistical Ties, and Why You Shouldn't Wear One by Andrew Vickers.

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.

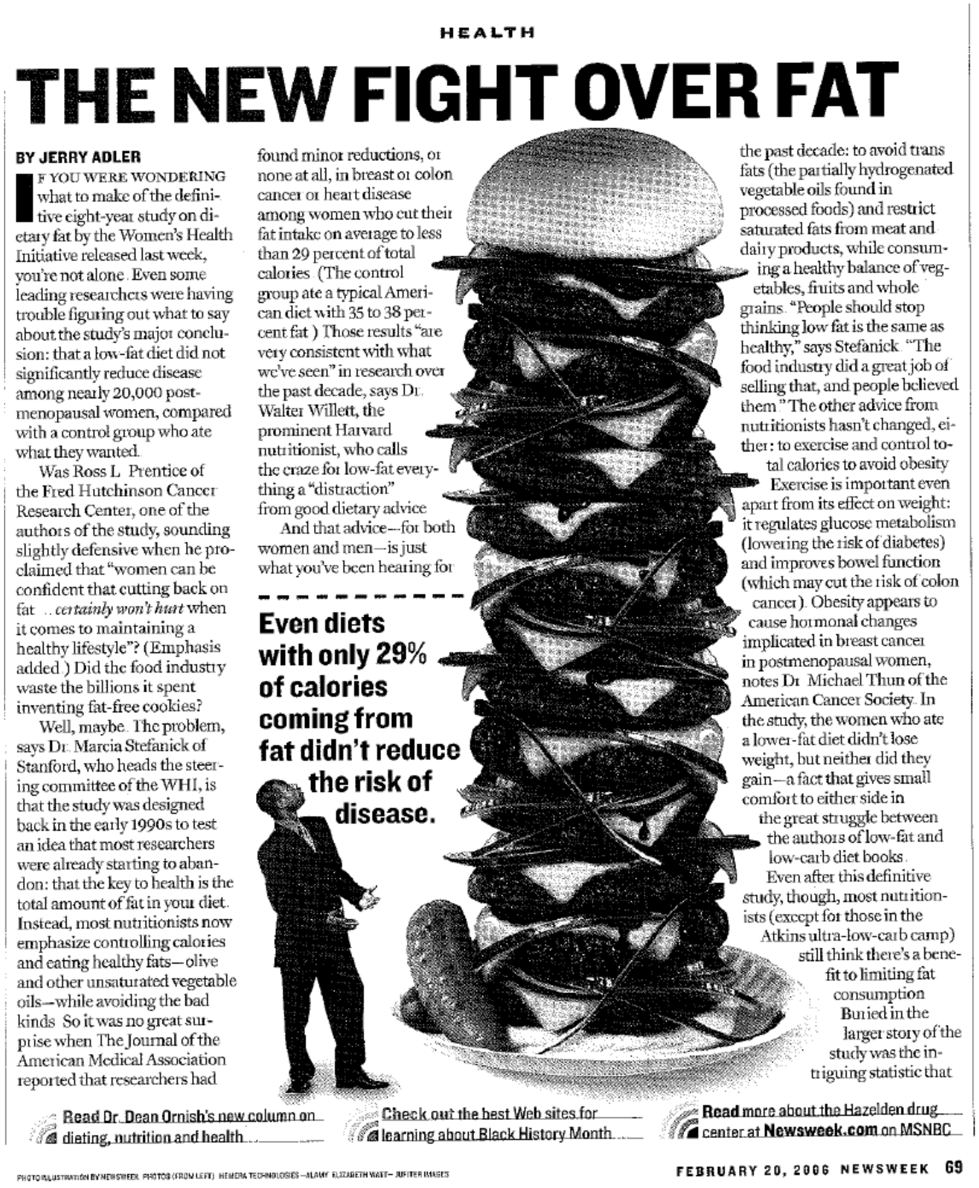
Growth restriction: RR 0.87 (0.66 – 1.16) Here, we have a big problem. The point estimate is a 13% reduction in the outcome, so the definitive statement that vitamins do not reduce this outcome is contradicted by the study's own data. Vitamins *did* appear to reduce this outcome, and the CI extends to 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 what may be the most important outcome has been interpreted as definitive evidence against effectiveness. If we knew that this observed estimate were correct, then vitamin supplementation, or at least further study, 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 (*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—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 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 negative studies, leading to Problem 1. The estimates and CI's should contribute to the interpretation, not just the p-value.

Think about these guidelines 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.

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 an example of a strong negative result that is well supported

NEJM, **354**: 1889-1900, 2006

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 may substantially 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 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.

Despite the careful attention to conflict of interest in medical research, this ubiquitious source of conflict receives little explicit attention.

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

Although there is a lot of pressure to make results seem as interesting as possible, 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).

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 solution is to always be vigilant.

There is one more 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 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 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 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. In some earlier years, for example, most written projects for this class did contain 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.

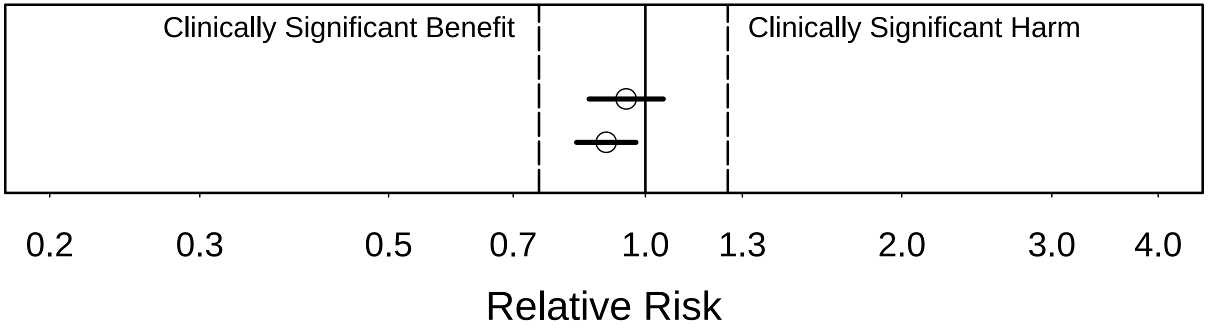
Check each sentence found and ask yourself

- 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?
- What if the lower confidence bound were true?

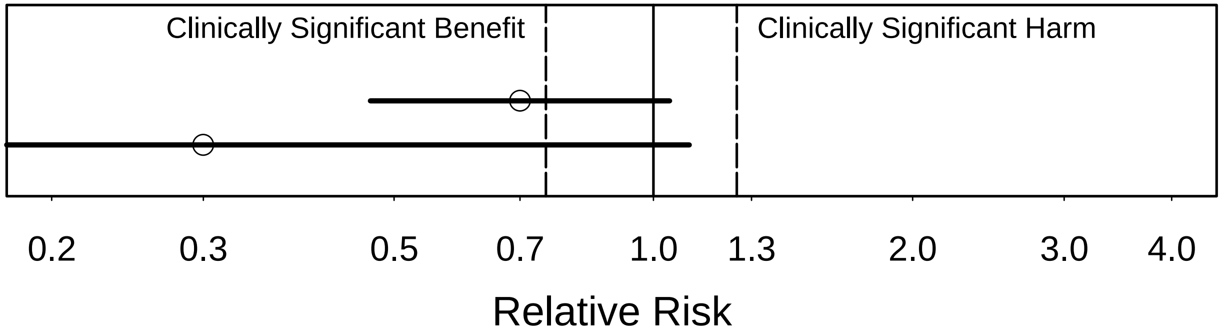
Additional searches: "failed", "lack", "absence", "disappeared", "only" Negative interpretations sometimes use these words, so you can also check them to make sure you didn't miss anything.

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 that reflect estimtates and CI's are available online.



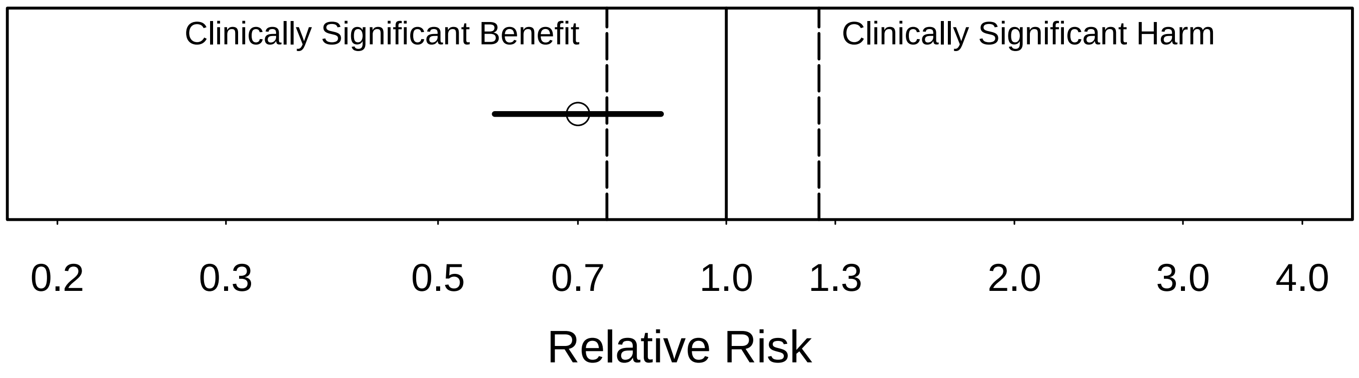
We found strong evidence against any substantial harm or benefit Because we have strong evidence against any values outside the CI, both these cases argue strongly that any effect is clinically unimportant. Note that this is true even though one is statistically significant.



Suggestion of substantial benefit The estimate would be an important benefit if true

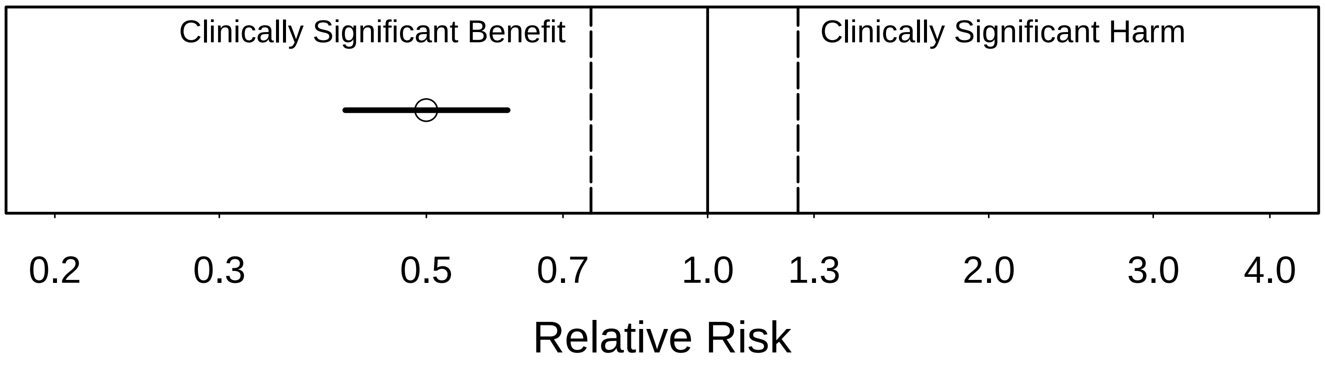
May be no effect (not statistically significant) The CI includes no effect

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.

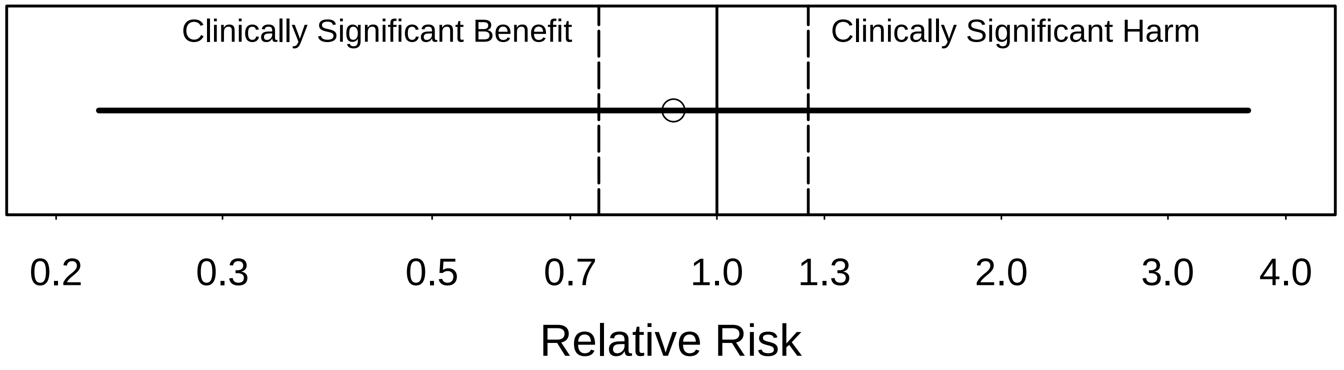


Strong evidence of benefit (statistically significant)

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.

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 vast 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 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," 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.