Common Biostatistical Problems And the Best Practices That Prevent Them

Biostatistics 209
April 21, 2009

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 findings

There may be exceptions

Please let me know about additions or disagreements

During lecture, or

Later (peter@biostat.ucsf.edu)

Problem 1. P-values for establishing negative conclusions

The P-value Fallacy:

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.

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

How about:

$p>0.05$ + Power Calculation = No effect

How about:

$p>0.05$ + Power Calculation = No effect

Still no good!

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

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 differences in sample size.

CONSORT guidelines: "There is little merit in calculating the statistical power once the results of the trial are known".

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

Additional references:

1958, D.R. Cox: "Power . . . is quite irrelevant in the actual analysis of data."

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.

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!

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

Example: Treatment of an acute infection

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

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

"No difference in death rates"
"No significant difference in death rates"
"No statistical difference in death rates"

Example: Treatment of an acute infection

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

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

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

NEJM, **354**: 1796-1806, 2006.

"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."

Preeclampsia: RR 1.20 (0.82 – 1.75)
Growth restriction: RR 0.87 (0.66 – 1.16)
Serious outcomes: RR 0.79 (0.61 – 1.02)

Women's Health Initiative study on fat consumption and breast cancer

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

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

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

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

Breast Cancer Mortality
HR 0.77 (0.48-1.22)

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

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

Often followed (but then ignored when interpreting)

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

The estimate is the value most supported by the data

The confidence interval includes values that are not too incompatible with the data

There is strong evidence against values outside the CI

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."

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?

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

“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.”

But then the paper probably won't end up in NEJM!

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.

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

Be vigilant (and be honest)!

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.

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 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."

Easy to slip into relying on “p>” reasoning

- Yes or No reasoning more natural
- Focus on p-values engrained in research culture
- Real level of uncertainty often inconveniently large, which can make results seem less interesting

Be vigilant

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

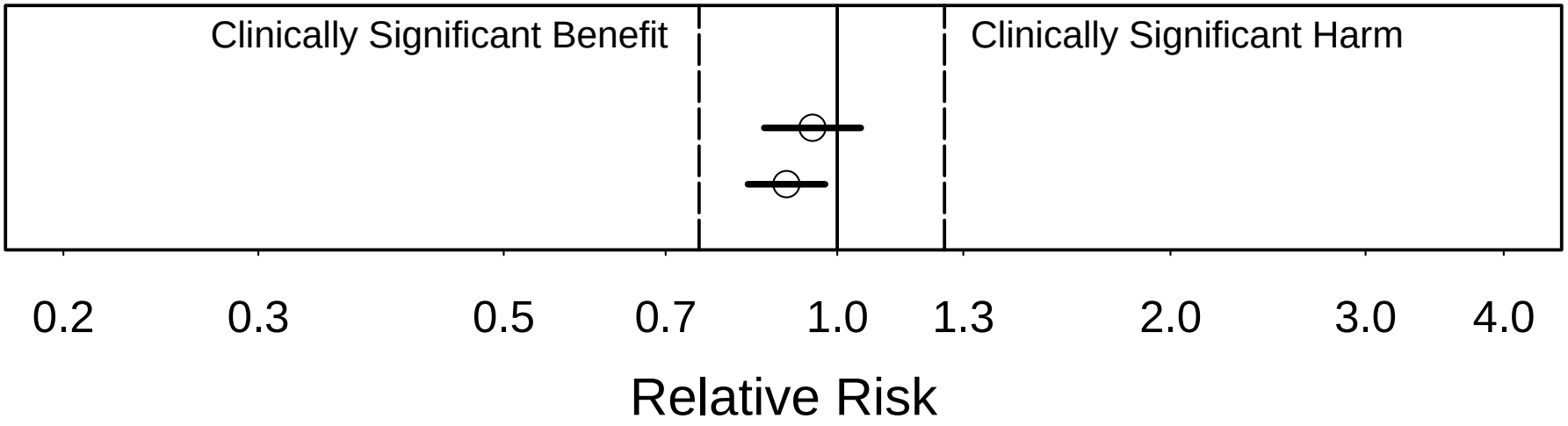
How to check negative interpretations:

Perform searches for words “no” and “not”

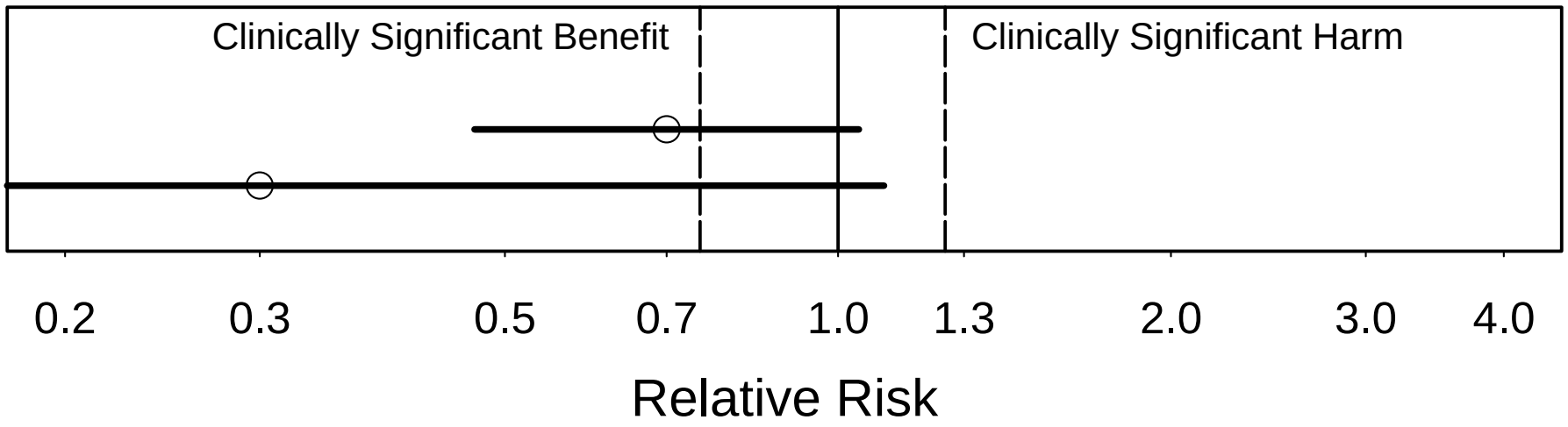
Check each sentence found

- 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?

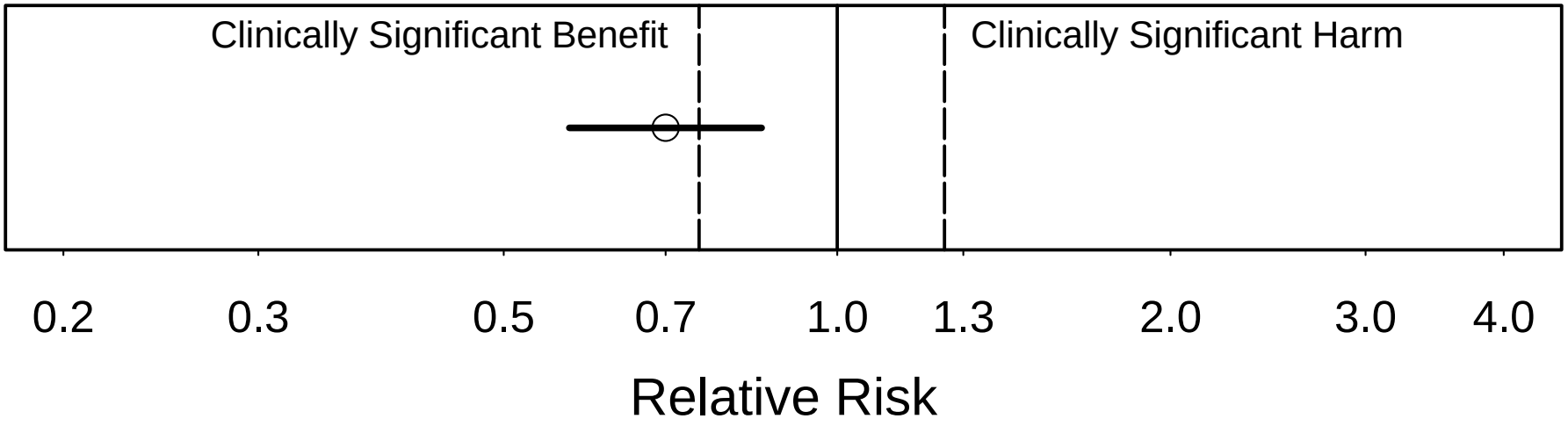
Additional searches: “failed”, “lack”, “absence”, “disappeared”, “only”



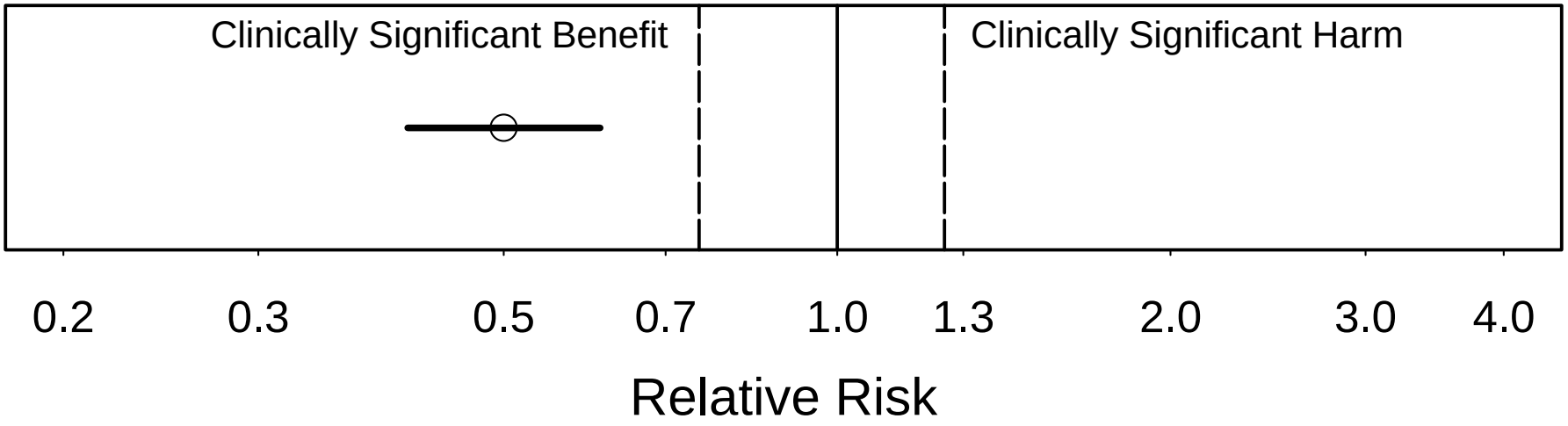
We found strong evidence against any substantial harm or benefit.



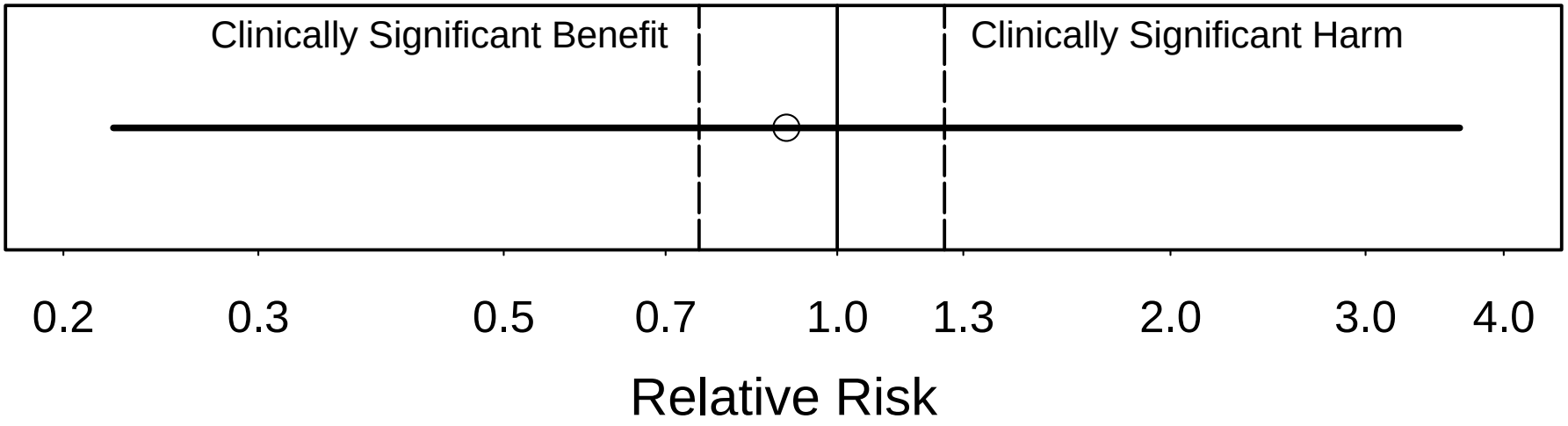
Suggestion of substantial benefit
May be no effect (not statistically significant)



Strong evidence of benefit (statistically significant)
Substantial benefit appears likely, but CI too wide to rule out clinically unimportant benefit



Strong evidence of substantial clinical benefit



No conclusions possible due to very wide CI

Also see online resource

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.”

Are large p-values good for anything?

“Due diligence” situations

Checking for possible assumption violations when little suspicion

Just need to state that you checked and nothing jumped out; don’t need to prove that nothing was present:

“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

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

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

Oddly, investigators often understate their conclusions using weak phrasing.

FRAM, nationwide study of fat abnormalities in HIV

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

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

Final: “We found evidence against any reciprocal increase in VAT in HIV-infected persons with peripheral lipoatrophy”

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)

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.”

Problem 3. Speculation about low power

"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."

What are they trying to say?

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

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

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

Problem 4. Exclusive reliance on intent-to-treat analysis

'Negative' study of vitamin E in diabetics (*JAMA* 2005)

"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.

WHI: Estimate of effect if adherent to low-fat diet:

Breast cancer HR 0.85 (0.71 – 1.02)

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.

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

- Interest in biological issues
- Double-blinded treatment

BP5b. Consider advanced methods to estimate causal effects.

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

Omnibus tests (like ANOVA)

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

Multiple comparisons adjustments

- each result detracts from the other

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
- some conditions that should differ

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

Reviewer's comment on a study examining effects of 4 different administration routes

"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**.

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.

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.”

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

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

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.)

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

BP7. Choose accuracy over conservatism whenever possible.

Problem 7. Entangled outcomes and predictors

Body mass index as a "predictor" of central fat

Many people have low peripheral and central fat
A few (both HIV and not) have low peripheral fat and high central fat

Low peripheral fat + low central fat → low BMI
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

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

Can only be treated if alive

Died after 2 days → max of 2 days treatment
Treated for 5 years → min of 5 years survival

Meaningless association

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

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

Technical problems

Unchecked assumptions
Ignoring dependence and clustering
Unclear details for time-to-event: operational definitions, early loss, event ascertainment
Missing data
Poor summaries (e.g., mean±SD for skewed data)
Showing inadequate or excessive precision
Poorly scaled predictors
Terms likely to be misread (“significant”)

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

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. doi:10.1371/journal.pone.0004149.

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.

Summary of Biostatistical Best Practices

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

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

- Interest in biological issues
- Double-blinded treatment

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

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 searches for words “no” and “not”

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”, “only”

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)

Take advantage of the faculty help that is available