

**Is Most Medical Research Wrong?
The Role Of Incentives And Statistical Significance.**

Jeffrey Hyman, DDS, PhD



Why Most Published Research Findings Are False.

Ioannidis JPA, 2005 PLoS Med 2(8): e124

Table 4. PPVs for different types of research.

$1 - \beta$	R	u	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.

RCT, randomized controlled trial.

DOI: 10.1371/journal.pmed.0020124.t004

We all know about the common problems in doing research.

Selecting study participants

- Selection bias
- Non-respondent bias:
- Volunteer or referral bias
- External validity
- Sampling bias
- Ascertainment Bias
- Prevalence-incidence bias .
- Berkson bias
- Healthy worker effect
- Detection bias: The risk factor investigated itself may lead to increased diagnostic
- Overmatching bias:

Information biases

- Recall Bias
- Reporting Bias
- Family information bias
- Measurement bias:
- Misclassification bias
- Reporting bias
- End-aversion bias
- Attention bias

Analysis

- Regression to the mean
- Competing death bias
- Publication bias
- Significance chasing bias:
- Confounding
- Residual confounding
- Interaction
- Expectation bias.
- Independence
- Compliance bias
- Withdrawal bias.

Part 1. Incentives and over-interpretation of results

**“I think I’ve been in the top 5% ... in understanding the power of incentives, and all my life I’ve underestimated it. And never a year passes but I get some surprise that pushes my limit a little farther.”
Charles Munger**

Incentives are rewards (that we want) and punishments (that we want to avoid)

If you want to know how people will behave, look at their incentives.



We are looking for the truth when we do research.

Is the search for truth our only reason for doing research?

Could we have any other incentives?

Financial

- We want to get grants
- We have a financial interest in the study
- We might want to support funding for a program
- We might want to continue funding for a program
- We want tenure
- We want a promotion

Psychology (ego)

- We think there is an association and we want to show it (confirmation bias)
- We want our studies to be published
- We want publicity
- We might have done work in this area before and we want to replicate it (confirmation bias)
- We have made an investment of time and money in doing the study and we want to show results
- We want people to think we are a good researcher
- We know about publication bias towards negative results

In many cases we need a publication in order to get the “reward” that we want or to avoid a “punishment”.

Publication bias and the need for an impressive result means we usually need to find a statistically significant result.

So in many cases the practical goal of our research is to find a $P < 0.05$

How does our strong desire for a $P < 0.05$ affect our results?

- We do extensive modeling with a range of variables. Then we **only** report the model with the most significant results (**selective reporting**) (**multiple comparisons**)
- We do extensive subgroup analyses (**multiple comparisons**)
- We compare extreme groups, such as the 1st and 5th quintiles.
- We use too large of a sample size for the effect we want to measure (such as national surveys)
- We use a 1 sided P value
- We don't try to publish papers with negative results
- We quickly do studies in hot fields
- We change study endpoints after looking at the data
- We investigate multiple associations between exposure and outcome (Pollock)
- We selectively cite the literature (**confirmation bias**)

Meta-analyses of 74 antidepressants studies submitted to the FDA showed that the increase in effect size ranged from 11 to 69% for individual drugs and was 32% overall. (Turner N Engl J Med 2008)

We do these things so pervasively that it has been reported that most non-null associations are inflated. (Ioannidis 2008)

Part 2. Problems Dealing with Statistical significance

-The meaning of P values

-The role of sample size and multiple comparisons

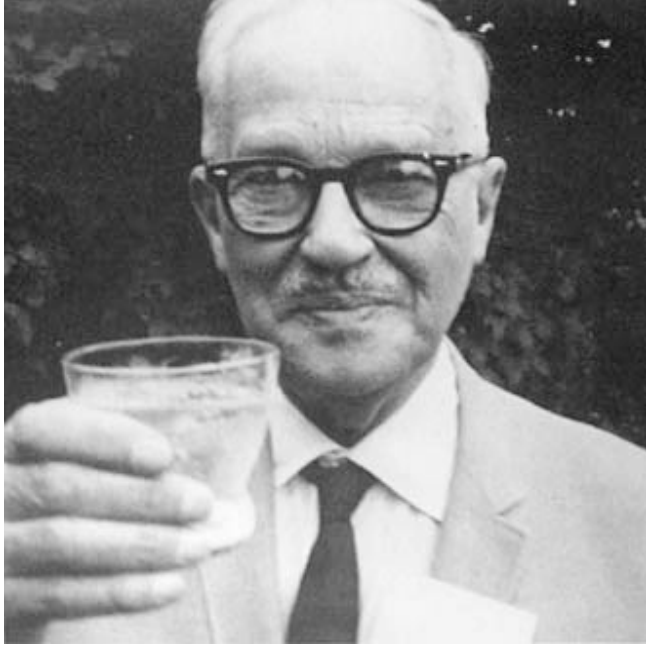
-Type I errors

P values and statistical significance are confusing.

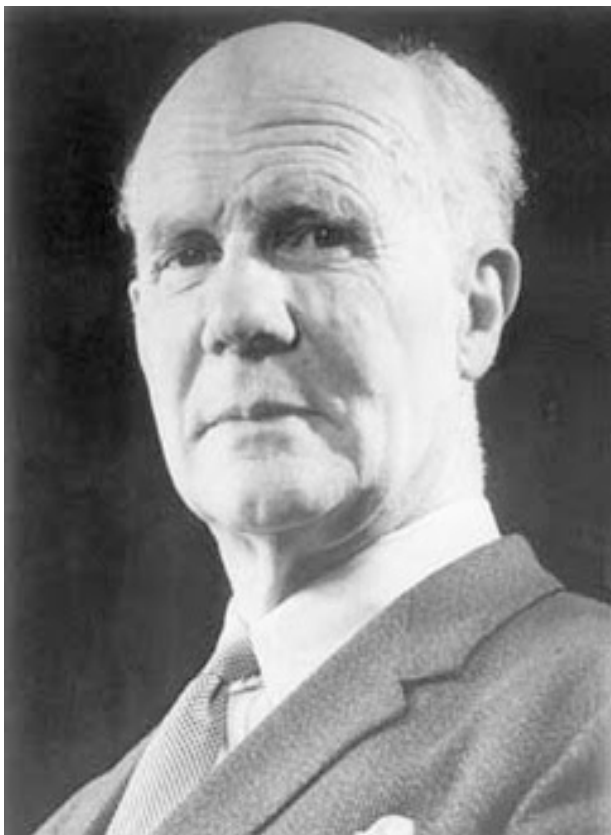


Sir Ronald Aylmer Fisher

He conceived the p value as a **flexible** inferential measure for judging the weight of evidence.



Jerzy Neyman



Egon Sharpe Pearson

Neyman-and Pearson embedded the P value into a test to formally reject the null hypothesis. **The hypothesis test was a decision rule, not an inference.**

What is the definition of a P value?

“The P value is the probability, **under the assumption of the null hypothesis H_0** , of obtaining a result **equal to or more extreme** than what was actually observed.”

Stated (hopefully) more simply:

If the two **populations** really have the **same mean**, the P value is the probability that random sampling would lead to a difference between sample means **as large or larger** than you actually saw.(Graphpad.com)

The claim has been made that P values actually **overstate** the evidence against the null hypothesis. (Goodman 1993)

Misconceptions about P Values.

People often think that the *P*-value is the probability that the null hypothesis is true given the data.

If $P=0.03$ there is a 97% chance that the difference you observed reflects a real difference between populations and a 3% chance that the difference is due to chance.

With a low *P*-value ($p < 0.001$), the findings must be true.

A nonsignificant difference (e.g., $P > .05$) means there is no difference between groups.

A statistically significant finding is clinically important.

Studies with *P* values on opposite sides of .05 are conflicting.

Studies with the same *P* value provide the same evidence against the null hypothesis.

A scientific conclusion or treatment policy should be based on whether or not the *P* value is significant.

$P = .05$ means that we have observed data that would occur only 5% of the time under the null hypothesis.

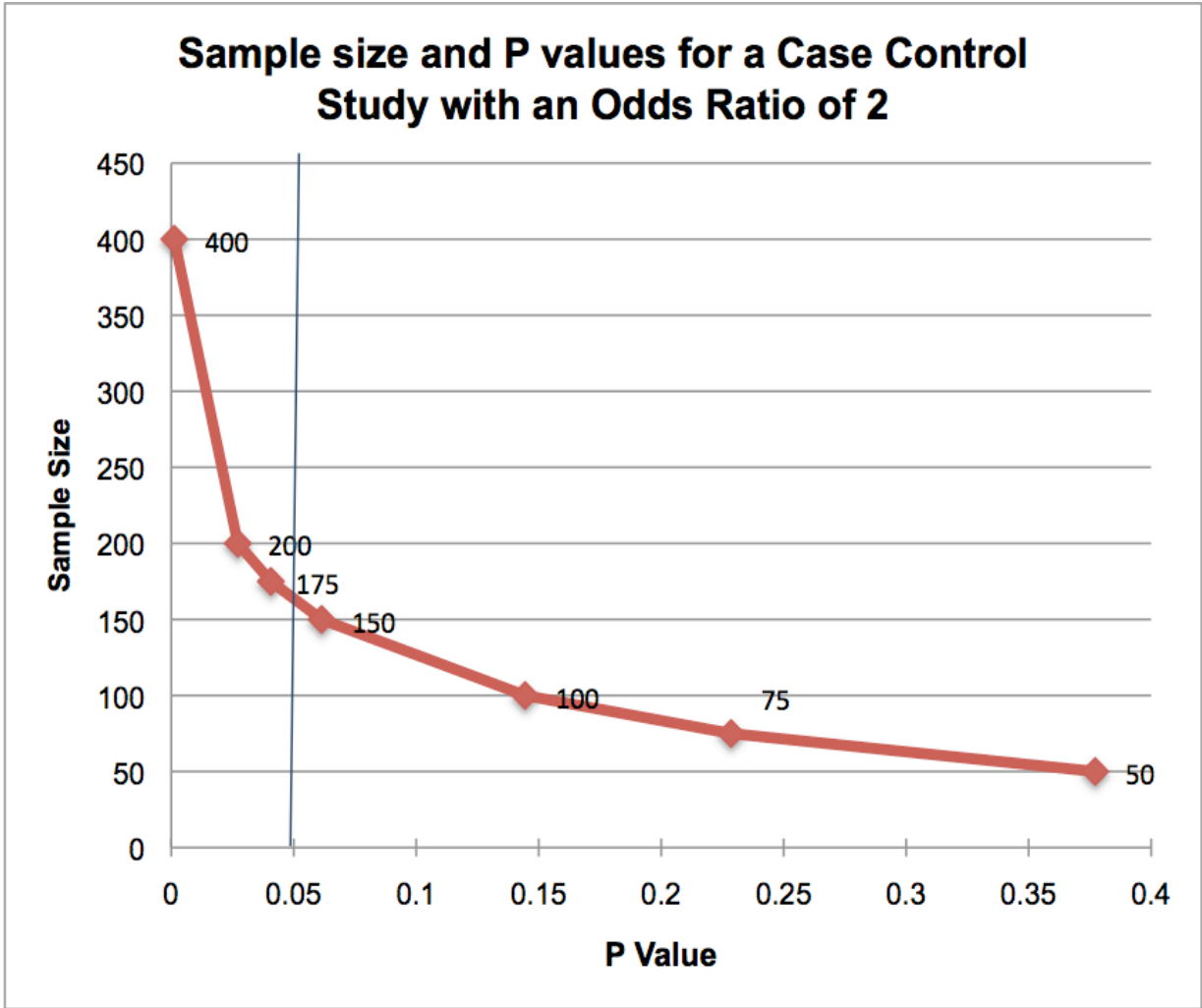
With a $P = .05$ significance level, the chance of a type I error will be 5%.

You should use a one-sided *P* value when you don't care about a result in one direction, or a difference in that direction is impossible.

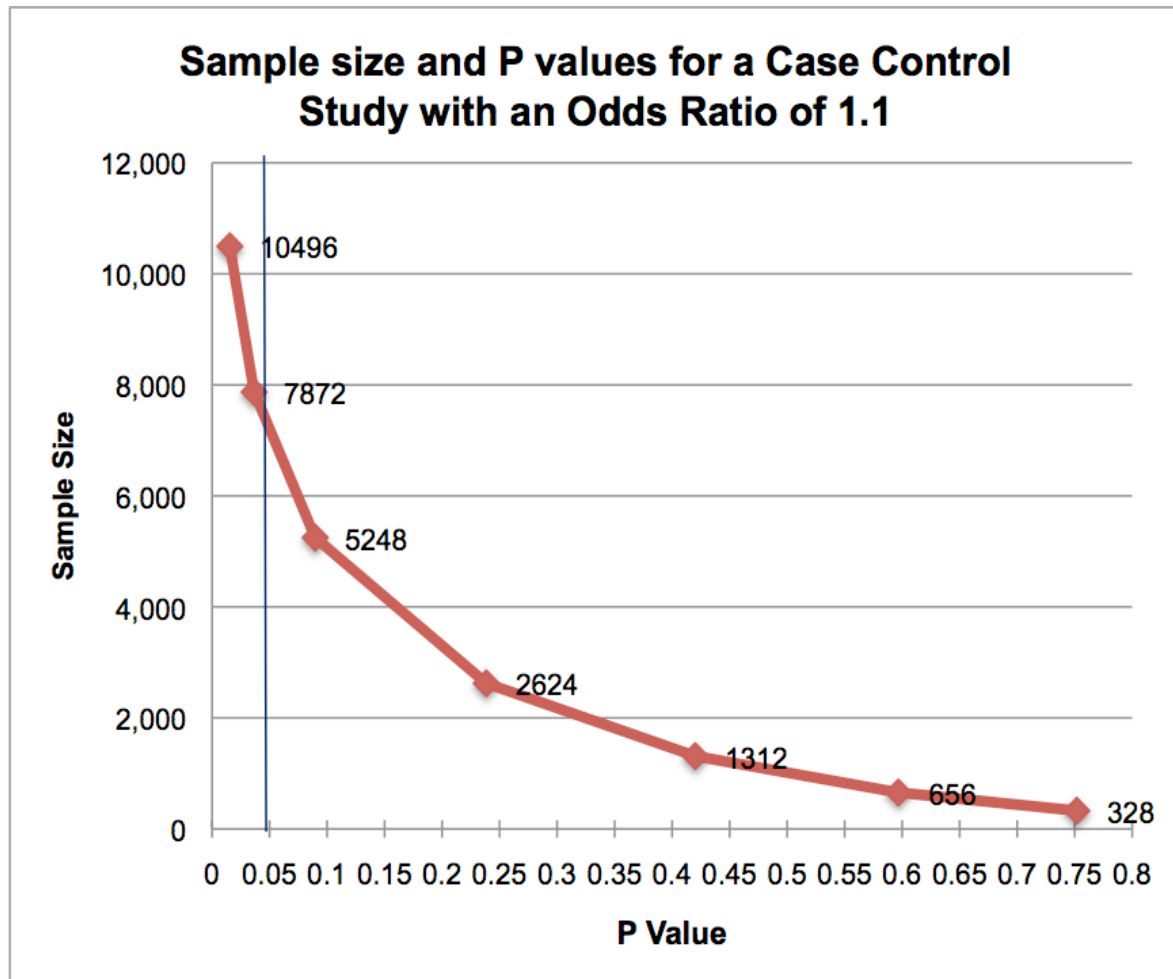
(Goodman, S Semin Hematol 2008)

The very important role of sample size in P values.

Too small of a sample.



Too large of a sample.

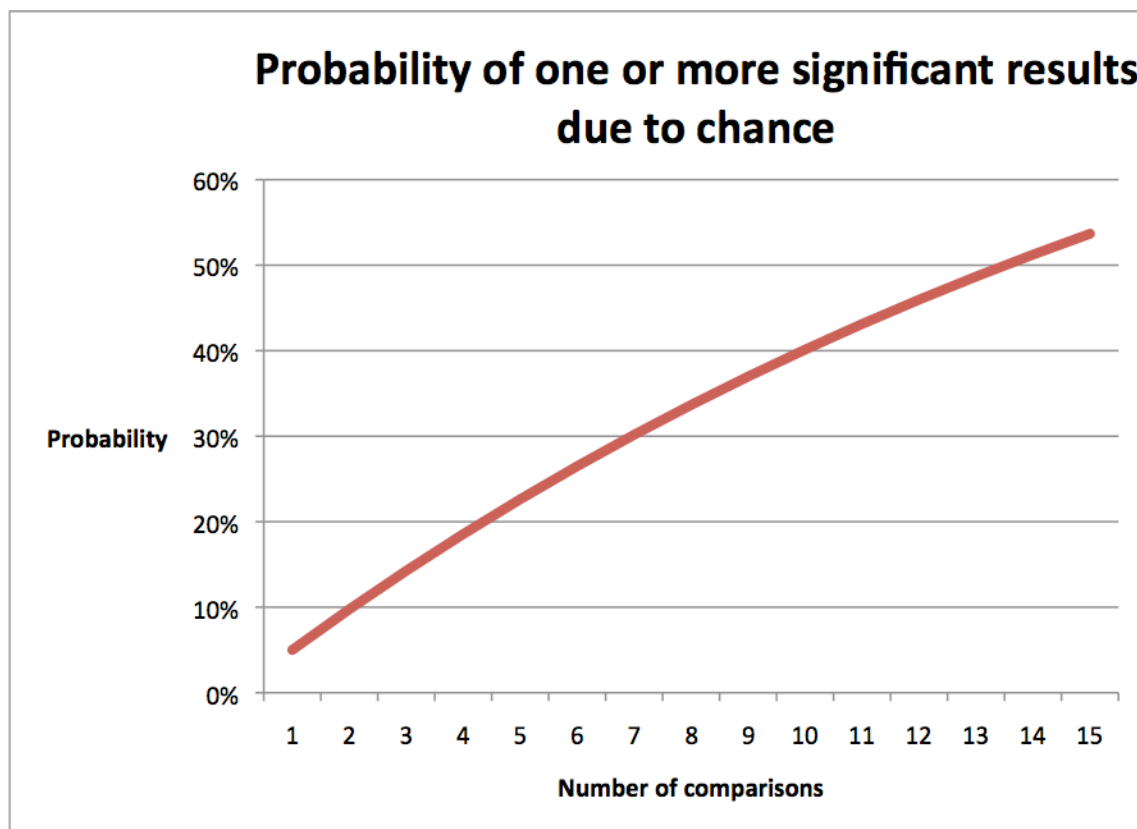


Multiple comparisons.

(doing lots of different models and subgroup analyses)

P values are based on making **one** comparison.

Multiple comparisons distort type 1 error rates (alpha), so the chance of a false positive is much larger than 5%.



With 100 multiple comparisons there is a 99% chance of getting a significant P value due to chance.

How should we interpret significant P values in these cases?

This was a discussion of type 1 errors in a single study. Is this the extent of type 1 errors?

Or

Can type 1 errors (false positives) propagate to the literature.

A hypothetical example.

- Researchers test 200 completely ineffective new drugs.
- About 10 trials out of the 200 will have a “significant” result due to chance.
- Only the 10 studies with significant results will be submitted for publication.
- **Five of these studies are published in major journals**

- Result: The type 1 error rate of each study was 5%, but the error rate in the literature is 100 percent (www.johndcook.com)

Could this really happen?

- The more prestigious a journal is the lower the percentage of papers that it accepts.
- Science accepts less than 8% of submissions.
- We know that doing research is difficult and most new hypothesis are wrong.
- Despite this, almost all published papers report positive results, so many of the studies in these journals could be wrong, despite being statistically significant.

Does this really happen?

- A study looked at some of the most **highly-cited** studies from the **most prestigious journals**.
- 32% were found to have either incorrect or exaggerated results.
- Of studies with a 0.05 p-value, 74% were incorrect.

(Ioannidis JAMA 2005)

Type 1 errors combined with publication bias

- 94% of published trials of 74 antidepressants studies submitted to the FDA were positive. Only 51% of the actual trials were positive. (Turner et al N Engl J Med 2008;)

- A study of 1,575 articles on cancer markers published in 2005 found that 95.8% reported positive results. (Kyzas 2007)

We have looked at the impact of incentives, type 1 errors, and problems with P values:

Can we predict which studies are more likely to be wrong?

- Small studies with significant results
- Studies with more flexible designs, outcome measures, and models
- Studies with significant results and a small effect measure (like odds ratio =1.1)
- The hotter the field and the more people doing research in it
- Studies where there are strong financial interests
- Studies with strong pre-existing beliefs by researchers
- The hotter the scientific field
- Fields with high throughputs such as microarrays

(Ioannidis 2005)

Problems with publication bias and selective reporting of results have also been identified in randomized controlled trials, which are our gold standard for clinical research. (Dwan K, et. al 2008)

Random sampling does not guarantee balanced or 'good' samples, especially in small studies.

What is a 'poor' random sample?

Example.

If you draw repeated **random** samples of size 100 and 1000 from a population with 50% women these are the largest and smallest number of women in the samples (Schoenbach 2009)

	N=100	N=1,000
Sample with the most women	68	54.9
Sample with the fewest women	33	45.0

Peer review does not solve these problems.

“... we know that the system of peer review is biased, unjust, unaccountable, incomplete, easily fixed, often insulting, usually ignorant, occasionally foolish, and frequently wrong.”

Richard Horton, FRCP FMedSci, editor-in-chief of *The Lancet*

- The agreement between reviewers is low.
- Reviews have been shown to miss many mistakes
- Reviewers can be biased against institutions, original work, and work that disagrees with what they have published
- Reviewers have been known to steal ideas.
- Peer review rejected 2 papers that led to Nobel prizes and the identification of B lymphocytes. (Smith 2010)

These type of problems lead to many expensive clinical trials having their results reversed.

Examples

- Two 1993 studies concluded that vitamin E prevents cardiovascular disease. That claim was overturned in 1996 and 2000.
- A 1996 study concluding that estrogen therapy reduces older women's risk of Alzheimer's was overturned in 2004.
- A major study concluded there's no evidence that statins help people with no history of heart disease. Cost of statins: more than \$20 billion per year, of which half may be unnecessary.
- A panel of the Institute of Medicine concluded that having a blood test for vitamin D is pointless: almost everyone has enough D for bone health. Cost of vitamin D: \$425 million per year.
- Numerous studies concluding that popular antidepressants work by altering brain chemistry have now been contradicted
- A *Lancet* paper that suggested that the MMR caused autism:
- A *New England Journal of Medicine* article that seemed to show that rofecoxib was safer than the traditional non-steroidal anti-inflammatory drugs.

(Begley) Newsweek Jan 21, 2011 and Smith (2010)

Causality and observational studies.



Sir Austin Bradford Hill

Hills Criteria of Causation.

- **Temporal Relationship:** Exposure always precedes the outcome.
- **Strength:** the size of the association as measured by appropriate statistical tests. The stronger the association, the more likely it is that the relation of "A" to "B" is causal.
- **Dose-Response Relationship:** If a dose-response relationship is present, it is strong evidence for a causal relationship.
- **Consistency:** The association is consistent when results are replicated in studies in different settings using different methods.
- **Plausibility:** The association agrees with currently accepted understanding of pathological processes.
- **Consideration of Alternate Explanations:** Determine the extent to which researchers have taken other possible explanations into account and have effectively ruled out such alternate explanations.
- **Experiment:** The condition can be altered (prevented or ameliorated) by an appropriate experimental regimen.
- **Specificity:** This is established when a single putative cause produces a specific effect. This is considered by some to be the weakest of all the criteria.
- **Coherence:** The association should be compatible with existing theory and knowledge. It is necessary to evaluate claims of causality within the context of the current state of knowledge within a given field and in related fields.

In an observational study we can not meet all the requirements. to establish causality; i.e. temporality, experiment, multiple studies.

In observational studies we should talk about associations, not causality.

Summary

- Be aware of how we overstate our results in an effort to get statistically significant results
- Be aware of the limitations of P values and statistical significance
- Don't over interpret significant results
 - Being significant does not make a results true or importance
 - Not being significant does not make a result falso
- Do power calculations for each study and be like Goldilocks - don't make your study too big or too small, make it just right
- Watch for problems like multiple comparisons, subgroup analysis, and selective reporting
- Be aware of the situations where study results are more likely to be wrong
- Remember the effects of publication bias
- In observational studies speak about associations, not causality