

## More Information: Common Statistical Pitfalls

- Over-emphasis on p-values
- Multiple comparisons/data dredging
- Residual confounding
- Overfitting
- Improper comparison group

## Over-Emphasis on P -Values 1

- Statistical significance does not guarantee clinical significance.
- Large studies have so much statistical precision (statistical power) that the tiniest differences will be "statistically significant."
- Example: a study of about 60,000 heart attack patients found that those admitted to the hospital on weekdays had a significantly longer hospital stay than those admitted to the hospital on weekends ( $p < .03$ ), but the magnitude of the difference was too small to be important: 7.4 days (weekday admits) vs. 7.2 days (weekend admits).
- Pay attention to effect size and confidence intervals.

## Over-Emphasis on P-Values 2

- On the flip side, lack of statistical significance is not proof of the absence of an effect.
- Studies may miss effects if they are insufficiently powered (lack precision).
- Example: A study of 36 postmenopausal women failed to find a significant relationship between hormone replacement therapy and prevention of vertebral fracture. The odds ratio and 95% CI were: 0.38 (0.12, 1.19), indicating a potentially meaningful clinical effect. Failure to find an effect may have been due to insufficient statistical power for this endpoint.
- Pay attention to effect size and confidence intervals.

Ref: Wimalawansa et al. *Am J Med* 1998, 104:219-226.

## Data Dredging/Multiple Comparisons

- "If you torture your data long enough they will confess to something"
- If you run 20 tests, you can expect 1 false positive (if your cut-off for statistical significance is 0.05).
- Ways to increase your type I error (chance of a false positive):
  - Run a lot of comparisons (statistical tests)
  - Do subgroup analyses
  - Adjust your definitions of "exposure" and "outcome"

Slide 5

## Data Dredging/Multiple Comparisons

- Example, subgroup analysis (a clever illustration of multiple comparisons!):
- In 1980, researchers at Duke randomized 1073 heart disease patients into two groups, but treated the groups equally. Not surprisingly, there was no difference in survival. But, when they divided the patients into 18 subgroups based on prognostic factors, they found a significant difference in survival between "group 1" and "group 2" in a subgroup of 397 patients (with three -vessel disease and an abnormal left ventricular contraction).
- *How could this be since there was no treatment?*
- The difference resulted from the combined effect of small imbalances in the subgroups. Tortured data!

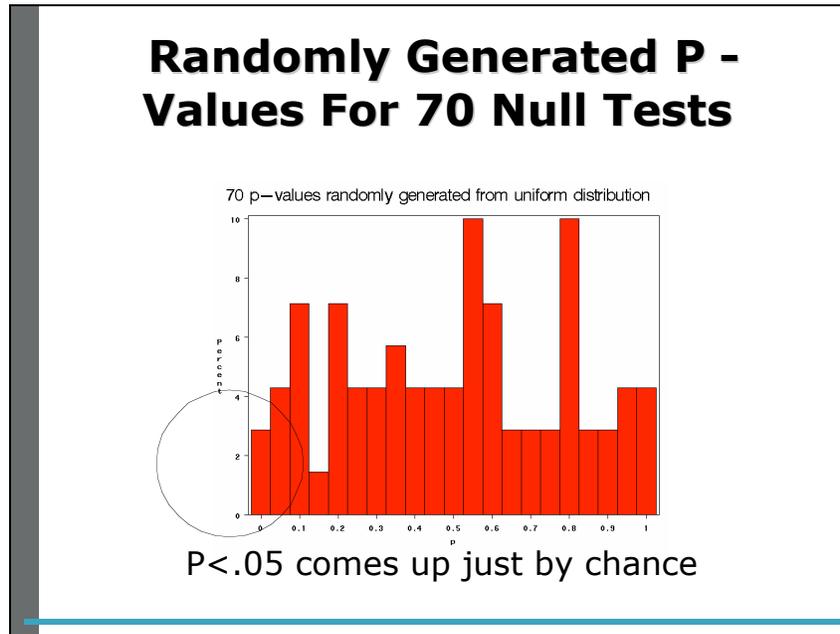
Ref: Lee et al. *Circulation*, 61: 508-515, 1980.

Slide 6

## Multiple Comparisons

- Hypothetical example, multiple statistical tests:
- Researchers compared nutrient intakes between injured and non-injured athletes. They measured athletes' nutrient intakes from both a food - frequency questionnaire (FFQ) and a 3-day food diary (70 variables total). They found three significant differences in nutrient intakes between injured and non-injured athletes (vitamin K and E as measured on the FFQ, and fat as measured by the food diary).
- But these findings were likely due to chance. With 70 comparisons, 3 false positives are expected.

Slide 7



Slide 8

- ### Multiple Comparisons
- Hypothetical example, tweaking the definitions of exposure and outcome:
  - Researchers performed a cross-sectional study to look for associations between alcohol intake and memory in the elderly. They tried several definitions of “moderate drinking” and “high memory performance” and found one significant association (1 drink per day and highest quintile on the test).
  - But chopping the data in many other ways did not yield significant results.

## Multiple Comparisons

- Ways to spot inflation of type I errors:
  - Authors did not distinguish between planned and exploratory comparisons
  - Authors did not distinguish between *a priori* and *ad hoc* subgroup analyses
  - Definitions of “exposure” and “outcome” appear arbitrary

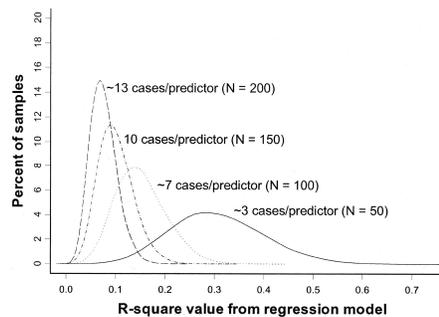
## Residual Confounding

- You cannot completely wipe out confounding simply by adjusting for variables in multiple regression unless variables are measured with zero error (which is usually impossible).
- Residual confounding can lead to significant adjusted odds ratios (ORs) as high as 1.5 to 2.0 if measurement error is high.
- Hypothetical Example: In a case-control study of lung cancer, researchers identified a link between alcohol drinking and cancer in smokers only. The OR was 1.3 for 1 - 2 drinks per day (compared with none) and 1.5 for 3+ drinks per day. Though the authors adjusted for number of cigarettes smoked per day in multivariate regression, we cannot rule out residual confounding by level of smoking (which may be tightly linked to alcohol drinking).
- Questions to ask yourself: Is the effect moderate in size? Are there strong confounders in play? Was the exposure, outcome, or strong confounder measured with considerable error/lack of precision?

## Overfitting

- In multivariate modeling, you can get highly significant but meaningless results if you put too many predictors in the model.
- The model is fit perfectly to the quirks of your particular sample, but has no predictive ability in a new sample.
- Example (hypothetical): In a randomized trial of an intervention to speed bone healing after fracture, researchers built a multivariate regression model to predict time to recovery in a subset of women (n=12). An automatic selection procedure came up with a model containing age, weight, use of oral contraceptives, and treatment status; the predictors were all highly significant and the model had a nearly perfect  $R^2$ -square of 99.5%.
- This is likely an example of overfitting. The researchers have fit a model to exactly their particular sample of data, but it will likely have no predictive ability in a new sample.
- Rule of thumb: You need at least 10 subjects for each additional predictor variable in the multivariate regression model.

## Overfitting



Pure noise variables still produce good  $R^2$  values if the model is overfitted. The distribution of  $R^2$  values from a series of simulated regression models containing only noise variables.

(Figure 1 from: Babyak, MA. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine* 66:411-421 (2004).)

## Wrong Statistical Comparison

- In RCTs or cohort studies with a control group, the relevant statistical comparison is: treatment/exposure group versus control group. Statistically significant improvement in the treatment group is not the relevant endpoint.
- Example: In a placebo-controlled randomized trial of DHA oil for eczema, researchers found a statistically significant improvement in the DHA group but not the placebo group. The abstract reports: "DHA, but not the control treatment, resulted in a significant clinical improvement of atopic eczema." However, the improvement in the treatment group was not significantly better than the improvement in the placebo group, so this is actually a null result.

Ref: Koch et al. *Br J Dermatol.* 2008 Apr;158(4):786-92.

## Wrong Comparison

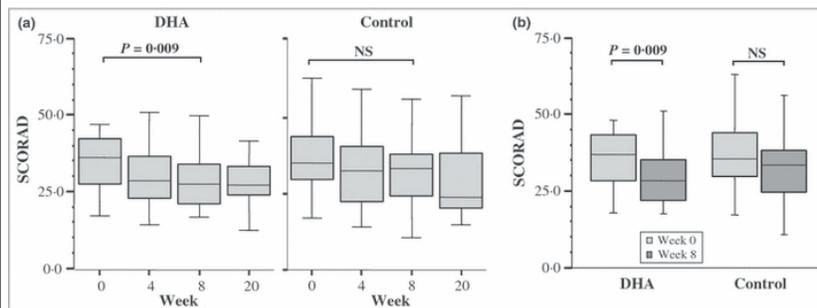


Figure 3 from: Koch C, Dölle S, Metzger M, Rasche C, Jungclas H, Rühl R, Renz H, Worm M. Docosahexaenoic acid (DHA) supplementation in atopic eczema: a randomized, double-blind, controlled trial. *Br J Dermatol.* 2008 Apr;158(4):786-92. Epub 2008 Jan 30.