Pitfalls of Biomedical Research

Prof. Oluwadiya K.S.
Ekiti State University
www.Oluwadiya.com

Damn, Damn Lies, and Statistics

-Mark Twain

Lies, Damned Lies, and Medical Science

MUCH OF WHAT MEDICAL RESEARCHERS CONCLUDE IN THEIR STUDIES IS MISLEADING, EXAGGERATED, OR FLAT-OUT WRONG. SO WHY ARE DOCTORS—TO A STRIKING EXTENT—STILL DRAWING UPON MISINFORMATION IN THEIR EVERYDAY PRACTICE? DR. JOHN IOANNIDIS HAS SPENT HIS CAREER CHALLENGING HIS PEERS BY EXPOSING THEIR BAD SCIENCE.

By David H. Freedman

Damn lies and medical science

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

factors that influence this problem and some corollaries thereof

is characteristic of the field and can vary a lot depending on whether the

Why most published research findings are false

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Investigator Paradigm Effect

Damn lies and medical science

ORIGINAL CONTRIBUTION

218 JAMA, July 13, 2005—Vol 294, No. 2 (Reprinted)

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

LINICAL RESEARCH ON IMPOR-

- 80% of non-randomized studies were wrong
- 25% of supposedly gold-standard randomized trials
- 10% of large randomized trials

Why then are errors so common in medical research?

- Mistakes and errors
- Deliberate fraud
  - Pressure
    - Publish or perish
    - Grants
    - Recognition and fame

Introduction

- The primary purpose of research is to conduct a scientific, or, scholarly investigation into a phenomenon, or to answer a burning question.
- Research may be defined as a systematic approach to problem solving.

What is a pitfall?

- A hidden danger or unsuspected difficulty (Webster)
- A pitfall is a conceptual error into which, because of its specious plausibility, people frequently and easily fall.
- It is "the taking of a false logical path“ that may lead the unwary to absurd conclusions, a hidden mistake capable of destroying the validity of an entire argument.

Steps in Performing Research

- **Research Problem** →What, When
- **Literature Review** →What, When, How, Why
- **Conceptual & Theoretical Frameworks** →What, Why
- **Variables & Hypotheses** →What, How
- **Research Design** →How
- **Population & sample** →Who, What
- **Data Collection** → How
- **Data Analysis** →Why
- **Results and findings**

Research Design Pitfalls

1. Not choosing the right study design
2. Not seeking the advice of a statistician on study design
3. Not specifying the hypotheses
4. Not specifying the outcome measures
5. Not anticipating potential confounders
6. Not specifying the randomization and blinding procedures

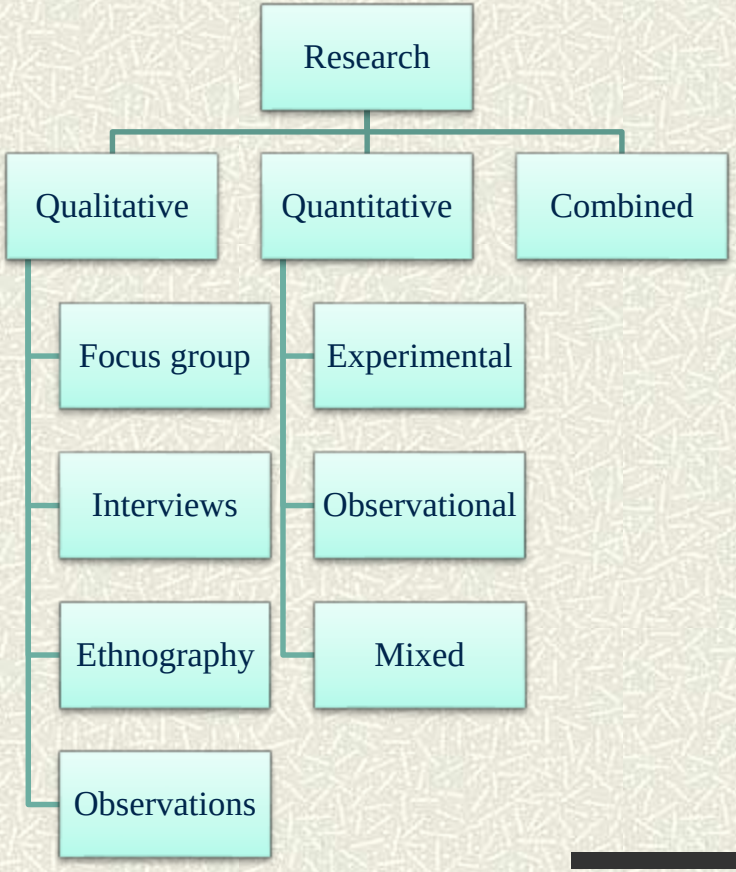
Research Design Pitfalls

(1) Not seeking statistical advice on study design

- May be very important
- What measurement levels should be used for each variable?
- Need for pilot study
- Priori hypothesis

Research Design Pitfalls

(2) Not choosing the right study design



Research Design Pitfalls

(3) Not specifying the priori hypotheses

1. Can lead to data dredging or multiple testing with its attendant errors (More on this later)

Research Design Pitfalls

(3) Not specifying the hypotheses (contd)

- A serious potential pitfall is present when investigators collect a large amount of data and have not pre-planned how they are to analyze the data. If an investigator is blessed with a abundance of data ... he can select those data which confirm his hypothesis that a relationship exists.
- The major problem here is that the investigator decides how the data are to be analyzed after he has "eyeballed" or studied the data. After the investigator has perused the data, he may decide to analyze only certain parts of the data while neglecting other parts. When the investigator has not planned the data analysis *beforehand*, he may find it difficult to avoid the pitfall of focusing only on the data which look promising (or which meet his expectations or desires) while neglecting data which do not seem "right" (which are incongruous with his assumptions, desires, or expectations).

Lipset, Trow, and Coleman (1970)

Research Design Pitfalls

Investigator Loose Procedure Effect

1. Not specifying the outcome measures
2. Not anticipating potential confounders
3. Not specifying the randomization and blinding procedures

Population and Sample Pitfalls

- Representative sampling is one of the most fundamental tenets of inferential statistics: the observed sample must be representative of the target population in order for inferences to be valid

Population and Sample Pitfalls

- Not calculating the correct sample size
  - Small (inadequate) samples
  - Overlarge samples

Population and Sample Pitfalls

- Using the wrong population
  - Hospital data
  - Employee data
  - Ignoring potential cofounders in the population
  - Wrong sampling method

Sampling Methods

Probability Sampling

- Simple random sampling
- Stratified random sampling
- Systematic sampling
- Cluster (area) sampling
- Multistage sampling

Non-Probability Sampling

- Deliberate (quota) sampling
- Convenience sampling
- Purposive sampling

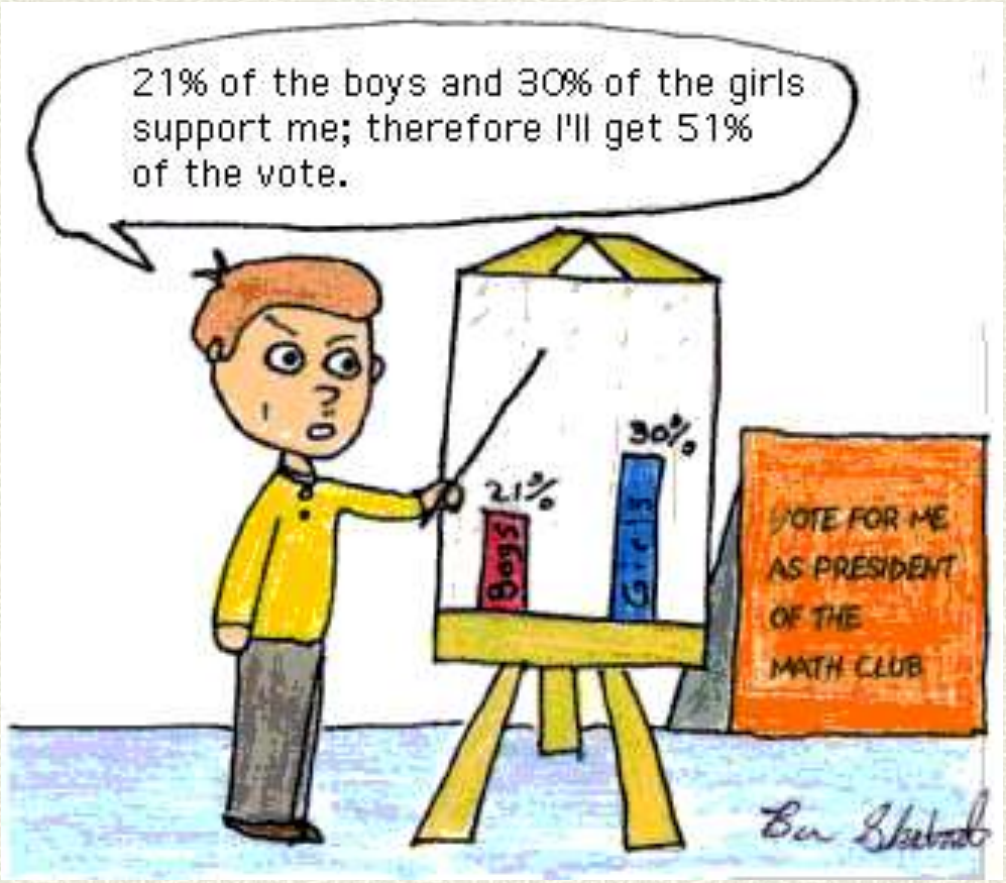
Data Collection Pitfalls

- Failure to follow the procedure as laid down in the methodology (*Experimenter Failure to Follow the Procedure Effect*)
- Poorly trained research personnel
- Poor supervisory procedure *(Investigator Loose Procedure Effect)*
- Poor supervision
- Outright fraud

Data Analysis Pitfalls

The Problem of Statistical Errors in research

- Widespread
- Long-standing
- Potentially serious
- Largely unknown
- Usually concerns basic, not advanced, statistics



Data Analysis Pitfalls

- Investigators at times fail to report that the data did not support their original hypothesis.
- Instead, after they have studied the data, they derive a new hypothesis that is supported by the data and then "verify" the new hypothesis by performing a statistical test on the same data from which it was derived
- *Although investigators may derive a new hypothesis from a completed study, the new hypothesis needs to be tested and verified in a subsequent study.*

(Lipset, Trow, & Coleman, 1970; Selvin, 1970).

Data Analysis Pitfalls

- Investigators at times collect incidental data that are not directly related to the hypotheses they are testing.
- If they fail to confirm their original hypotheses, they then perform a large number of statistical tests on the remaining data and report whatever significant results are obtained as "findings.“
- *This can easily lead to misleading conclusions*

Data Analysis Pitfalls

- Failure to report negative results.
- Investigators may discard all data of an experiment as bad data if not in agreement with theory, and start over
- The problem here is that if the investigator obtains positive results in a later study and publishes them without mentioning his earlier negative results, the reader is likely to conclude wrongly that the positive results are more stable, more easily replicable, or more valid than is actually the case

Data Analysis Pitfalls

- When an investigator obtains negative results that fail to confirm his hypothesis he is likely to check for computational errors in the data analysis or to run another data analysis
- However, when the original analysis confirms the investigators' hypothesis, it is unlikely that he will check for computational errors or run another analysis

(Friedlander, 1964).

Data Analysis Pitfalls

Using descriptive statistics incorrectly

- Use the mean and standard deviation ONLY to report normally distributed data: "Mean (SD) height was 72 cm (4.3 cm)."
- Use the median and interquartile range to report non normally distributed data: "Median (IQR) length was 9 cm (6 to 25 cm)."

Data Analysis Pitfalls

Using descriptive statistics incorrectly

- The shape of the distribution (normal or skewed) may determine the class of statistical test used to analyze the data ("parametric" or "nonparametric," respectively).
- Most biological data are not normally distributed; the median and IQR should be used in such situations

Data Analysis Pitfalls

Over-emphasis on p-values

- Statistical significance does not guarantee clinical significance.
  - 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).

Ref: Kostis et al. *N Engl J Med* 2007;356:1099-109.

Data Analysis Pitfalls

Over-emphasis on p-values

- Statistical significance does not guarantee clinical significance.
  - Clinically unimportant effects may be statistically significant if a study is large (and therefore, has a small standard error and extreme precision).

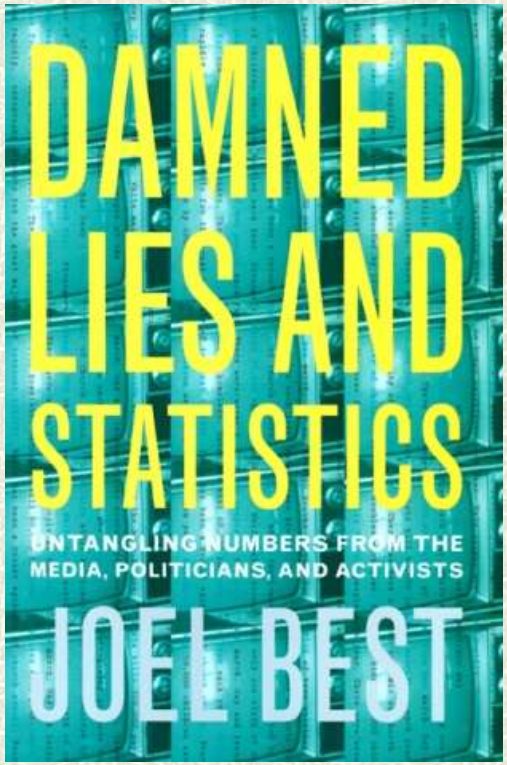
Pay attention to effect sizes and confidence intervals

Data Analysis Pitfalls

Over-emphasis on p-values

- Statistical significance does not imply a cause-effect relationship.

Interpret results in the context of the study design.



Myths about significance

- **Myth 1**: “If a result is not significant, it proves there is no effect.”
- **Myth 2**: “The obtained significance level indicates the reliability of the research finding.”
- **Myth 3**: “The significance level tells you how big or important an effect is.”
- **Myth 4**: “If an effect is statistically significant, it must be clinically significant.”

Problems with relying on *p*-values

- When sample size is low, *p* is usually too big
  - → if effect size is big, try bigger sample
- When sample size is very big, p can easily be very small even for tiny effects
  - e.g., mean IQ of men is 0.8pts higher than IQ of women, in a sample of 10,000: statistically significant, but is it clinically significant?
- When many tests are run, one of them is bound to turn up significant by random chance
  - → multiple comparisons: inflated Type-I errors

Multiple Testing or Data Dredging

- Performance of 2 or more related hypothesis tests using the same data set
- Examples:
  i. Suppose we consider the efficacy of a drug in terms of the reduction of any one of a number of disease symptoms. As more symptoms are considered, it becomes more likely that the drug will appear to be an improvement over existing drugs in terms of at least one symptom.

Wikipedia

Multiple Testing or Data Dredging

- Examples (Contd):

2. Suppose we consider the safety of a drug in terms of the occurrences of different types of side effects. As more types of side effects are considered, it becomes more likely that the new drug will appear to be less safe than existing drugs in terms of at least one side effect.

Wikipedia

Multiple Testing or Data Dredging

- Danger of Multiple Testing
  - For a single test, with significant level at 0.05 means that there is only a 5 percent chance that it is a spurious finding resulting solely from chance variations.
  - For two tests: the probability that at least one such analysis will yield a spurious, significant finding is greater than 5 percent.

    To determine the new probability level:

    The probability that a significant result would not be obtained in either of the two = .95 X .95 = 0.902
    Subtract this from 1.
    1-.902 = .098.

Multiple Testing or Data Dredging

- Danger of Multiple Testing
  - For 10 tests:
    The probability that a significant result will not be obtained in any of the ten tests is $(0.95)^{10}$
    $=0.59$
    The new probability 1-0.59
    $=0.41$
  - Formula is: $1-(1-n)^x$
    n=Significant level
    x=number of independent tests

Multiple Testing or Data Dredging

- Danger of Multiple Testing: Real life Example
  - 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.
  - Then they divided the patients into 18 subgroups based on prognostic factors.
  - In a subgroup of 397 patients (with three-vessel disease and an abnormal left ventricular contraction) survival of those in "group 1" was significantly different from survival of those in "group 2" ($p<.025$).
  - ***How could this be since there was no treatment?***

(Lee et al. "Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease," *Circulation*, 61: 508-515, 1980.)

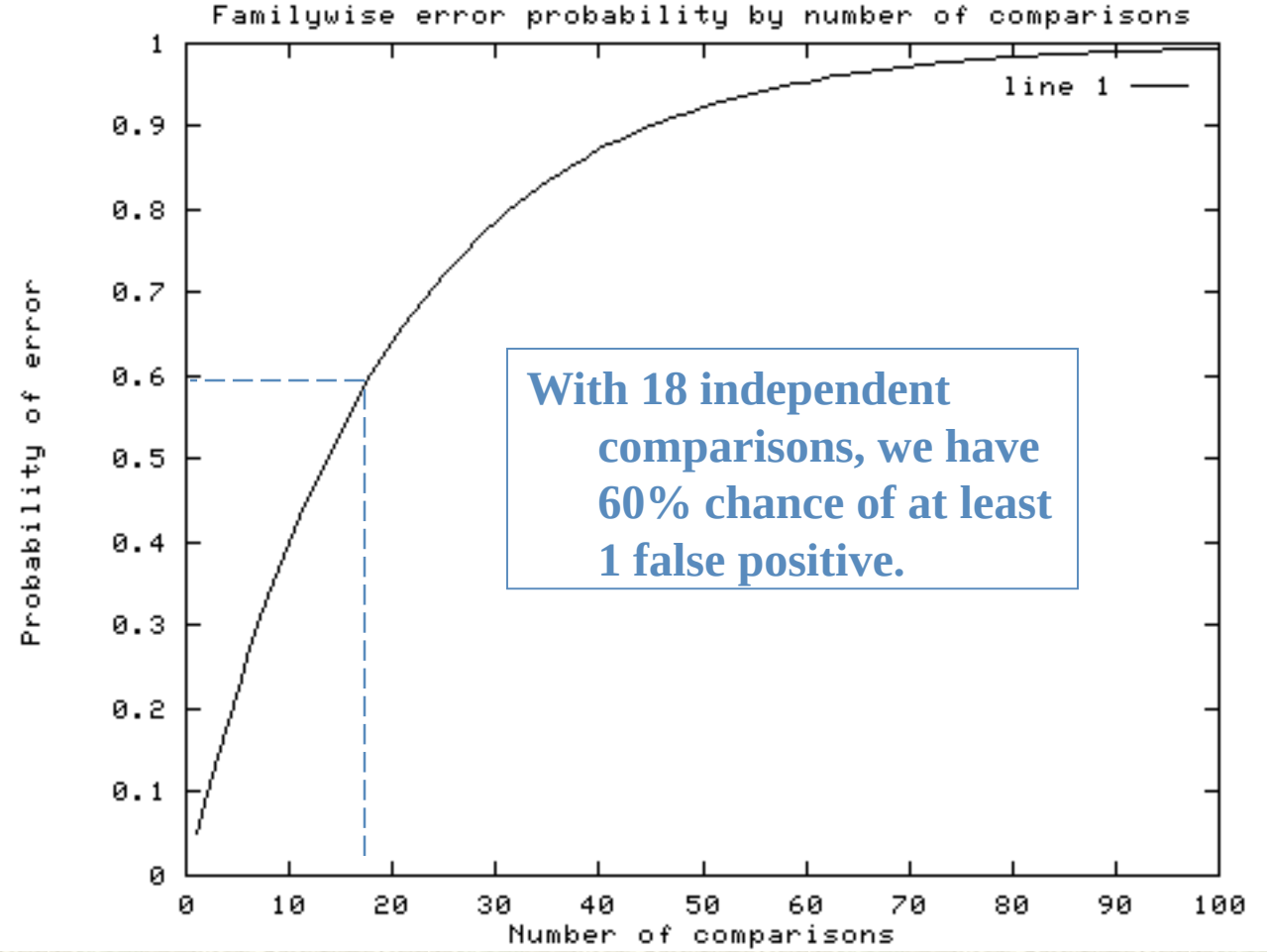
Multiple Testing or Data Dredging

- Danger of Multiple Testing: Real life Example

  - If we compare survival of "treatment" and "control" within each of 18 subgroups, that's 18 comparisons.
  - If these comparisons were independent, the chance of at least one false positive would be…

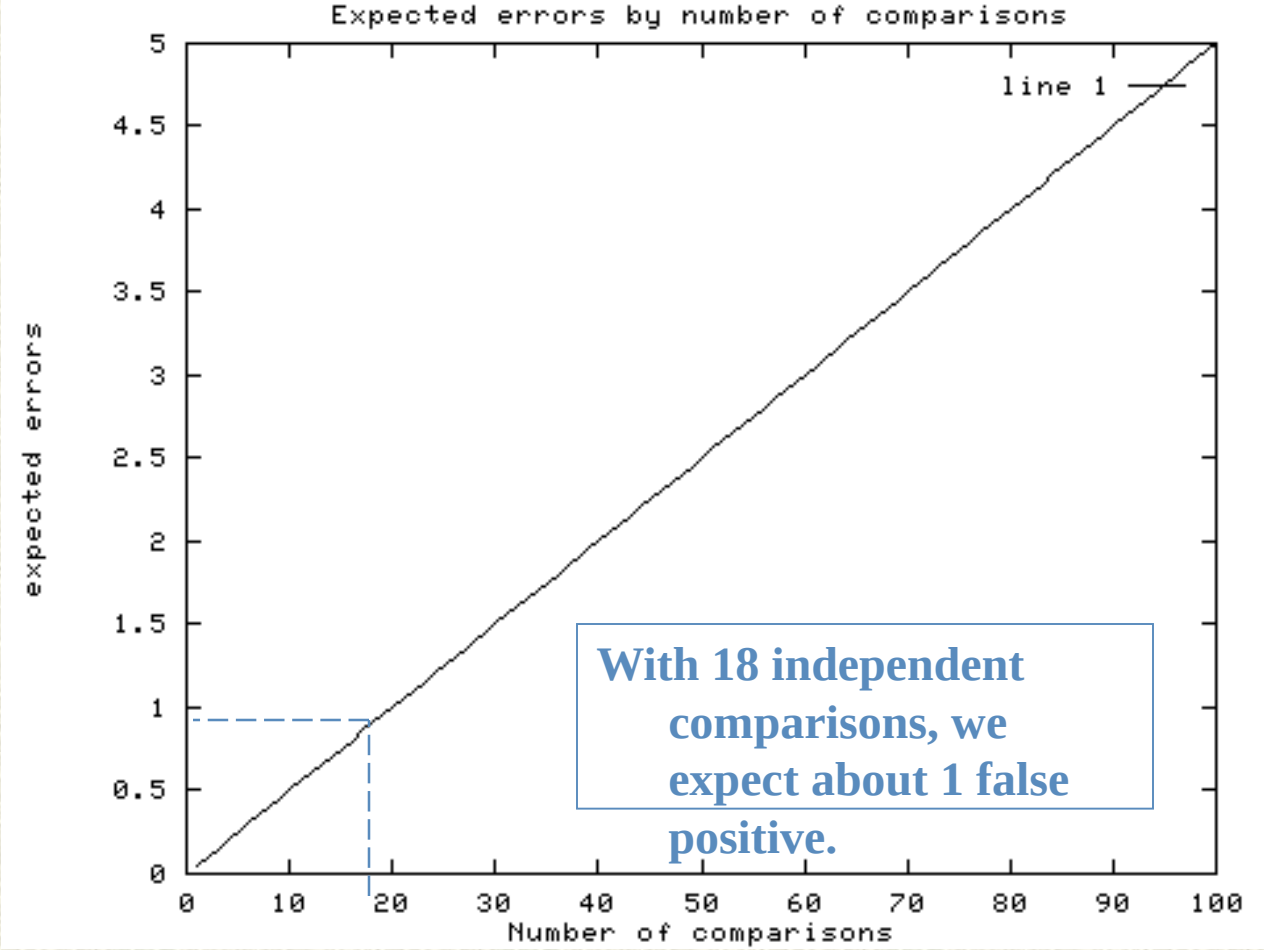
$$1-(.95)^{18}=.60$$

Multiple testing



With 18 independent comparisons, we have 60% chance of at least 1 false positive.

Multiple testing



Sources of multiple testing

Source	Example
Multiple outcomes	a cohort study looking at the incidence of breast cancer, colon cancer, and lung cancer
Multiple predictors	an observational study with 40 dietary predictors or a trial with 4 randomization groups
Subgroup analyses	a randomized trial that tests the efficacy of an intervention in 20 subgroups based on prognostic factors
Multiple definitions for the exposures and outcomes	an observational study where the data analyst tests multiple different definitions for "moderate drinking" (e.g., 5 drinks per week, 1 drink per day, 1-2 drinks per day, etc.)
Multiple time points for the outcome (repeated measures)	a study where a walking test is administered at 1 months, 3 months, 6 months, and 1 year
Multiple looks at the data during sequential interim monitoring	a 2-year randomized trial where the efficacy of the treatment is evaluated by a Data Safety and Monitoring Board at 6 months, 1 year, and 18 months

In the medical literature…

- Hypothetical example:
  - Researchers wanted to compare nutrient intakes between women who had fractures and women who had no fractures.
  - They used a food-frequency questionnaire and a food diary to capture food intake.
  - From these two instruments, they calculated daily intakes of all the vitamins, minerals, macronutrients, antioxidants, etc.
  - Then they compared fracturers to non-fracturers on all nutrients from both questionnaires.
  - They found a statistically significant difference in vitamin K between the two groups ($p<.05$).
  - They had a lovely explanation of the role of vitamin K in injury repair, bone, clotting, etc.

In the medical literature…

- Hypothetical example:
  - What's going on? Almost certainly artifactal (false positive!).

Factors indicative of chance findings

1. **Analyses are exploratory.**	The authors have mined the data for associations rather than testing a limited number of *a priori* hypotheses.
2. **Many tests have been performed, but only a few p-values are "significant".**	If there are no associations present, .05*k significant p-values (p<.05) are expected to arise just by chance, where k is the number of tests run.
3. **The "significant" p-values are modest in size.**	The closer a p-value is to .05, the more likely it is a chance finding. According to one estimate*, about 1 in 2 p-values <.05 is a false positive, 1 in 6 p-values <.01 is a false positive, and 1 in 56 p-values <.0001 is a false positive.
4. **The pattern of effect sizes is inconsistent.**	If the same association has been evaluated in multiple ways, an inconsistent pattern of effect sizes (e.g., risk ratios both above and below 1) is indicative of chance.
5. **The p-values are not adjusted for multiple comparisons**	Adjustment for multiple comparisons can help control the study-wide false positive rate.

*Sterne JA and Smith GD. Sifting through the evidence—what's wrong with significance tests? *BMJ* 2001; 322: 226-31.

Errors

- Type I error
  - Claiming a difference between two samples when in fact there is none.
  - Also called α error.
  - Typically 0.05 is used

Errors

- Type II error
  - Claiming there is no difference between two samples when in fact there is.
  - Also called β error.
  - The probability of not making a Type II error is $1 - \beta$, which is also called the *power* of the test.
  - It usually cannot be detected without a proper power analysis

Errors

Test Result

Truth	No difference H_0	Shows Difference H_1
No difference H_0	No Error	Type I α
Shows Difference H_1	Type II β	No Error

How to Increase the Power?

i.e. reduce type II error

1. Increase the size number
2. Reduce variation between measurements
3. The effect of intervention should be stronger

Sample Size Calculation

- Also called "*power analysis*".
- When designing a study, one needs to determine how large a study is needed.
- Power is the ability of a study to avoid a Type II error.
- Sample size calculation yields the number of study subjects needed, given a certain desired power to detect a difference and a certain level of P value that will be considered significant.
  - Many studies are completed without proper estimate of appropriate study size.
  - This may lead to an erroneous "negative" study outcome.

Sample Size Calculation

- Depends on:
  - Level of Type I error: 0.05 typical
  - Level of Type II error: 0.20 typical
  - One sided vs. two sided: nearly always two-sided
  - Inherent variability of population
    - Usually estimated from preliminary data
  - The difference that would be meaningful between the two assessment arms.

Keep In Mind That

- No study is perfect
- All data is dirty is some way or another; research is what you do with that dirty data
- Measurement involves making choices

Be Critical About Numbers

- Every statistic is a way of summarizing complex information into relatively simple numbers.
- How did the researchers arrive at these numbers?
- Who produced the numbers and what is their bias?
- How can key terms be defined & in how many different ways?

Be Critical About Numbers

- How was the choice for the measurement made?
- What type of sample was gathered & how does that affect result?
- Is the statistical result interpreted correctly?
- If comparisons are made, are they appropriate?
- Are there competing statistics?

Be Critical About Numbers

With one foot in a bucket of ice water, and one foot in a bucket of boiling water, you are, on the average, comfortable.

Thanks for your attention