Session 112 PD - Lying with Statistics: Pitfalls in Medical/Drug Testing

Moderator:
Russell A. Osborn, FSA, CCERA, MAAA

Presenters:
Russell A. Osborn, FSA, CCERA, MAAA
Ronald L. Wasserstein, Ph.D.
Lying with Statistics: 
Pitfalls in Medical/Drug Testing

RUSS OSBORN, FSA, CFA, CERA
VP, Risk Methodology, Aegon/Transamerica
October 17, 2017
Issues with Published Studies

Why Most Published Research Findings Are False
John P. A. Ioannidis

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 false relationships among the factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings
Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there
Finding Patterns in Clouds

In a given scientific field:

• The smaller the # studies conducted
• The smaller the effect sizes
• The greater the number of tested relationships
• The lesser the (pre-specified) selection of relationships
• The greater the flexibility in designs, definitions, outcomes, and analytical modes
• The greater the financial and other interests and prejudices
• The “hotter” the scientific field (with more scientific teams involved)

More likely that the research findings are false
Design Flaws

• Parachutes don’t work
• To a man with a hammer, everything looks like a nail
• Lead-time bias
• Younger people live longer than older people
• Healthy volunteer bias
Tinkering

• Look at that shadow on the wall
• Confound it!
• Mis-over-correcting
• Dredging data
• Dredging data sets
Publication Bias

• Move along: Nothing to see here
• No admission for the decent
• Happy sponsors
Interpretation

• Fun with correlations
• Misdirection in the conclusion
• Tunnel vision
• Convenient selection of statistics
Parachutes Don’t Work

• Test a wide range of parachute sizes, ranging from 3 inches to 21 inches
• Conclusion?
• Example:
  • Much quoted Tulane study, purporting to show that a low-carb diet was not only more effective at losing weight than a “low”-fat diet, but also was better for your heart
  • 30% calories from fat vs. 35% (also with many confounding variables)
  • Much lower % fat consistently shows the opposite result
Man with a Hammer

• a.k.a. Over-Detection
• Example: tumor detection
• False positives → readily “cured”
Lead-time Bias

(a) Lead-time bias. Early detection necessarily advances the date of diagnosis of a cancer compared with clinical detection; however, in this case, although the individual lives longer with a diagnosis of cancer, there is no change in the date of death.

(b) Length-biased sampling. The arrows represent tumors. The body of the arrow represents the preclinical growth period; the arrowhead represents the onset of symptomatic disease. The vertical dotted lines represent application of a screening test. Screening is more effective at detecting slow-growing, less aggressive tumors, because these lesions have a longer preclinical period during which they can be detected by tests.

Younger People Live Longer

• 5-year survival rates
• Early detection and treatment ➔ Higher survival at the end of 5 years!
• Better measure: Change in Quality-Adjusted Life Years
Shadows on the Wall

• Focus on the results of a marker rather than the objective itself
• Example: cholesterol is a marker – not an objective
• Be careful interpreting studies that focus on the marker rather than on longevity or QALY
Confound It!

- Cause (independent variable)
- Effect/outcome (dependent variable)
- Other factor (confounding variable)
Confounding Variables: Classic Example
Over-correction

Before

After
Mis-over-correcting

• Classic example:

  • Tobacco lobbyist argues that infant mortality data needs to be adjusted for the confounding variable of birth weight.

  • Depending on how the adjustment is performed and how infant mortality is defined (given that smoking during pregnancy contributes to both lower birth weight and infant mortality), the adjustment can fully remove the observed mortality effect, or sometimes even arrive at the opposite conclusion (exploiting residual noise in the data).

  • Q.E.D.
Move Along: Nothing to See Here

• Negative results rarely published
• But there is bound to be a four-leaf clover somewhere: \( p \times N > x \) if \( N \) is larger enough!

• Possible solutions:
  • Registration of studies (both empirical and observational)
  • Summary reports of all such studies
  • Replication journal dedicated to debunking bad results (and credence given to publications in such a journal from academic tenure committees)
  • More focus on meta-studies / meta-analysis
    • Must be done wisely!
No Admission for the Decent

• No Randomized Double-Blind Placebo Control (RDBPC) study, then go home
  • Even if it would likely harm/kill someone?
Happy Sponsors

- Who wrote the check?
Fun with Correlations

![Internet Explorer vs Murder Rate](chart1)

![Correlation of Changes](chart2)

![Scatter Plots](chart3)
Misdirection in the Conclusion

• Beware conclusions which mine the results for findings that were not part of the original objective or hypothesis (worse yet: which do not conform to the results!)
Tunnel Vision

• Over-reliance on a single statistic
  • Especially the case today for p-statistic
  • In fact, this topic has become so important, ...

• Worse yet: Biased selection of favorable statistics, after the fact
The Gold Standard

• Posit a plausible relationship (backed by a plausible, conceptual, causal model), state it as a hypothesis, and then test it (not the other way around!)
• Define success and failure in advance
• Register the study and report the results, regardless of the outcome
• Look at all statistics
• Look at all available studies
  • Both experimental and observational
  • Both clinical and epidemiological
Doctor, It Hurts When I $p$

RONALD L. WASSERSTEIN
Executive Director, American Statistical Association
October 17, 2017
The Talk

• They think they know all about it already, because they learned about it from others like them.
• It is not nearly as interesting as they thought it would be.
• They’ve stopped listening before you’ve stopped talking.
• Chances are, they now understand it even less.
Does “screen time” affect sleep habits of school age children?

Interactive vs passive screen time and nighttime sleep duration among school-aged children

Jennifer Yland, BA Candidate, Stanford Guan, MPH, Erin Emanuele, MPH, Lauren Hale, PhD
Received: February 17, 2015; Received in revised form: June 22, 2015; Accepted: June 24, 2015; Published Online: August 13, 2015

DOI: http://dx.doi.org/10.1016/j.sleh.2015.06.007
The researchers had hypotheses, based on previous research

- use of any form of electronic media negatively associated with sleep duration
- the strength of the association would vary based on the level of interactivity of the screen type
- interactive forms of screen time would be associated with shorter bedtime sleep duration compared to passive forms of screen time
What would these look like, if true?

• The more kids watch TV, play video games, or chat on line, the less sleep they get

• Kids who play video games or chat on line for X amount of time (various values of X) would sleep less than kids who just watch TV for that amount of time
Why were they interested?

• Lack of sleep (insufficient sleep duration) increases risk of poor academic performance as well as certain adverse health outcomes

• Is there a relationship between weekday nighttime sleep duration and screen exposure (television, chatting, video games)?
Who were the subjects?

- Nine-year-olds
- Selected from an ethnically diverse national birth cohort study, the Fragile Families and Child Wellbeing Study
- Data reported by both the child (n = 3269) and by the child's primary caregiver (n = 2770)
Fragile Families and Child Wellbeing Study

• Followed approximately 5000 children, born between 1998 and 2000, since birth.

• Data were collected in 20 cities with populations of at least 200,000 across the United States.

• The sample was designed to include a high number of unmarried parents and racial minorities, along with a high proportion of low socioeconomic status.”
<table>
<thead>
<tr>
<th></th>
<th>NONE</th>
<th>HALF AN HOUR OR LESS PER WEEKDAY</th>
<th>MORE THAN HALF AN HOUR BUT LESS THAN AN HOUR PER WEEKDAY</th>
<th>1-2 HOURS PER WEEKDAY</th>
<th>MORE THAN 2 HOURS PER WEEKDAY</th>
<th>REF</th>
<th>DK</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1A.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>D1B.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>D1C.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>D1D.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>D1E.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>D1F.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>D1G.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>D1H.</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
</tbody>
</table>

Hang out with friends? Do you spend no time at all, spend half an hour or less per weekday, more than half an hour but less than an hour per weekday, 1-2 hours per weekday, or more than 2 hours per weekday? 

Hang out with family members? 

Do household chores or help at home? 

Spend time on the computer doing school work? 

Spend time on the computer chatting or instant messaging with friends? 

Spend time on the computer or TV playing computer games? 

Spend time watching TV and movies? 

Attend practice or lessons or an after-school Program?
112 How many hours of **sleep** a night does {CHILD} usually get during the week?

ENTER NUMBER OF HOURS A NIGHT

OR

REFUSED........................................................................... -1

DON'T KNOW..................................................................... -2
13. Now think for a moment about a typical weekday for your family, including daytime and evening hours. How much time would you say {CHILD} spends watching television or watching videos on TV, either in your home or somewhere else?

IF LESS THAN 1 HOUR PER WEEKDAY, CODE AS ZERO.

PROBE: Do not count time {he/she} spends playing video games on TV.

_______
ENTER HOURS PER WEEKDAY
OR
REFUSED..............................................................-1
DON'T KNOW..........................................................-2
What MIGHT the researchers have found?

• TV/chatting/video games positively/negatively correlated with sleep duration (6 different outcomes)
• Sleep time (TV) > Sleep time (VG) > Sleep time (Chat) (six different possible orders)
• Child/caregiver reported (2 outcomes)
• We’re already up to 24 possible results, and we aren’t really done
What did the researchers find?

Three things
What did the researchers find?

Children who watched more than 2 hours/day of TV had shorter sleep duration compared with those who watched less than 2 hours/day (P<.001) by about 11 minutes.
What did the researchers find?

Children who spent more than 2 hours per day of chatting on the computer had shorter sleep duration than those who chatted less than 2 hours/day (P<.05) by about 16 minutes.
What did the NOT researchers find?

no significant association between playing videogames/working on the computer for more than 2 hours per day and weekday nighttime sleep duration
(over)simplified version of results

• TV watching over two hours/day reduces sleep duration by 11 minutes (on average)
• Chatting over two hours/day reduces it by 16 minutes
• But... video games have no significant impact on sleep duration
• Also: This amount of sleep loss is less than that reported in other studies
This is a fairly typical type of study

• Typical scientifically
• Typical statistically
• Atypical communication
Unfortunately, it makes all-too-typical mistakes
To understand these mistakes, let’s describe the null hypothesis significance testing procedure (NHSTP).
What is the null hypothesis significance testing procedure?

• Question(s) posed
• Data collected
What is the null hypothesis significance testing procedure (NHSTP)?

• Evidence from the data regarding the research question is summarized in a specific way:
  • Compute a “statistic” that measures the question of interest.
  • Compute the probability that statistic would be as “large” as it is or even larger UNDER THE ASSUMPTION that there is no effect (in this case, of TV watching on sleep duration).
  • This assumption of no effect is called the “null hypotheses.”
  • The probability computed is called the “p-value.”
What is the null hypothesis significance testing procedure (NHSTP)?

• If the p-value is “small enough,” the researcher concludes there is a “significant effect.”

• “Small enough” has come very commonly to mean $P < .05$. 
Certain assumptions must be made to compute a p-value

• An underlying statistical model
• Many things related to that model (randomness, representativeness, missing data, and so on)
• The null hypothesis
In terms of this example:

• The null hypothesis (informally stated) is that sleep duration is not associated with screen time (of various types).

• That is, when we calculate the p-value, we assume the answer to our question is no, there is no association between screen time and sleep duration.

• The p-value is calculated based on the assumption that there is no effect.

• The p-value is calculated based on the assumption that there is no effect.
What’s the logic?

• If the p-value is small, this means that it is relatively unlikely that we would have seen the data we saw if all the assumptions were true.

• So, we either had bad luck (random error), or one or more of the assumptions may not be true.

• One of those assumptions, the assumption of no effect, is commonly THE assumption that is thought to be untrue.
In the example:

- Children who watched more than 2 hours/day of TV had shorter sleep duration compared with those who watched less than 2 hours/day (P<.001) by about 11 minutes.
In the example:

This means that, if all of the assumptions are correct, including the null hypothesis, there is less than a 1 in 1000 chance that the researchers would have observed the result they did or one even larger. (The result they observed is an average difference of about 11 minutes from one group to the other.)
In the example:

- A 1 in 1000 chance is not very likely
- So it is not likely that, if all of the assumptions are correct, we would have observed the outcome we observed (11 minutes difference in sleep time) or one even larger.
- Therefore, we should evaluate these assumptions, including the null hypothesis
R.A. Fisher called such results “significant”
To Fisher, this meant that the result was worth further scrutiny

• Unfortunately, the word “significant” is loaded with meaning

• Statisticians and others draw the distinction between “statistical significance” and “practical significance”
What people tend to conclude in these situations? (What will the blogs say?)

- Research shows that children who watch TV more during the weekday sleep less than those who don’t.
- And from there it is a short walk to “TV is not good for kids and should be limited” or “TV is causing poor performance in school because it makes kids sleep less.”
- Authors’ conclusion in abstract: “No specific type or use of screen time resulted in significantly shorter sleep duration than another, suggesting that caution should be advised against excessive use of all screens.” – In other words, though not demonstrated in the study, all screen usage is suspect.
Starting a land war in Asia

The paper makes some other “classic blunders” as well
What is scientifically appropriate to conclude?

• The children in this study who watched more than 2 hours/day of TV had shorter sleep duration compared with those who watched less than 2 hours/day by about 11 minutes.

• If all of our assumptions, including those about the representativeness of the sample, are correct, the study suggests that nine year old children from this population who watch more than 2 hours/day of TV....
In the sleep research, even if all of our assumptions are correct...

• Does 11 minutes less sleep really matter? Why?
• Furthermore, the “11 minutes” measure is an estimate that has variance – we learn nothing about that variance from the way the data summary is reported (i.e., via a p-value)
And what if THIS had happened:

• Suppose the study showed that children who watched 2 or more hours of TV slept on average 90 minutes per night less than those who did not, but the p-value was 0.09.

• Is this result “insignificant”?
ASA statement articulates six principles

1. *p*-values can indicate how incompatible the data are with a specified statistical model.

2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

4. Proper inference requires full reporting and transparency

5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.
Biggest takeaway message from the ASA statement

Bright line thinking is bad for science
“(S)cienists have embraced and even avidly pursued meaningless differences solely because they are statistically significant, and have ignored important effects because they failed to pass the screen of statistical significance...It is a safe bet that people have suffered or died because scientists (and editors, regulators, journalists and others) have used significance tests to interpret results, and have consequently failed to identify the most beneficial courses of action.” (Rothman)
p equal or nearly equal to 0.06

- almost significant
- almost attained significance
- almost significant tendency
- almost became significant
- almost but not quite significant
- almost statistically significant
- almost reached statistical significance
- just barely below the level of significance
- just beyond significance
- "... surely, God loves the .06 nearly as much as the .05."
  (Rosnell and Rosenthal 1989)
p equal or nearly equal to 0.08

• a certain trend toward significance
• a definite trend
• a slight tendency toward significance
• a strong trend toward significance
• a trend close to significance
• an expected trend
• approached our criteria of significance
• approaching borderline significance
• approaching, although not reaching, significance
And, God forbid, p close to but not less than 0.05

- hovered at nearly a significant level (p=0.058)
- hovers on the brink of significance (p=0.055)
- just about significant (p=0.051)
- just above the margin of significance (p=0.053)
- just at the conventional level of significance (p=0.05001)
- just barely statistically significant (p=0.054)
- just borderline significant (p=0.058)
- just escaped significance (p=0.057)
- just failed significance (p=0.057)
Thanks to Matthew Hankins for these quotes

• https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/
A fundamental problem

We want $P(H|D)$ but p-values give $P(D|H)$
The problem illustrated (Carver 1978)

What is the probability of obtaining a dead person (D) given that the person was hanged (H); that is, in symbol form, what is $p(D|H)$?

Obviously, it will be very high, perhaps .97 or higher.
The problem illustrated (Carver 1978)

Now, let us reverse the question: What is the probability that a person has been hanged (H) given that the person is dead (D); that is, what is \( p(H|D) \)?

This time the probability will undoubtedly be very low, perhaps .01 or lower.
The problem illustrated (Carver 1978)

No one would be likely to make the mistake of substituting the first estimate (.97) for the second (.01); that is, to accept .97 as the probability that a person has been hanged given that the person is dead.

Inference is hard work.

- Simplistic (“cookbook”) rules and procedures are not a substitute for this hard work.
- Cookbook + artificial threshold for significance = appearance of objectivity
In a world where $p<0.05$ carried no meaning...

• What would you have to do to get your paper published, your research grant funded, your drug approved, your policy or business recommendation accepted?
You’d have to be convincing!
You will also have to be transparent
You could also

• List all planned comparisons in advance
• Adjust the “alpha level” (the p-value threshold) to account for the number of comparisons
• Focus more on effect sizes and confidence intervals
• Use multiple methods of analysis, and focus on results that are robust to the methods chosen
• Choose a really small p-value threshold
Wrapping up...
P-values themselves are not the problem, but...

• They are hard to explain
• They are easy to misunderstand
• They don’t directly address the question of interest
• When mixed with bright line thinking, they lead to bad science.
Haiku

Little p-value
what are you trying to say
of significance?

-Steve Ziliak
ron@amstat.org

@RonWasserstein