Pitfalls of P-hacking
Guest Lecture
Neuro 602

Paramita Saha Chaudhuri, PhD
Biostatistics
McGill University
paramita.sahachaudhuri@mcgill.ca

Who am I?

- Assistant Professor in Biostatistics (2014)
- Research interest in prediction modeling accuracy and privacy-preserving data analysis
- Collaborations in Kidney transplant, occupational asthma, HIV-AIDS, onco-nursing, neuro
- Teaching: introductory biostatistics, longitudinal modeling and prediction modeling; OH workshop

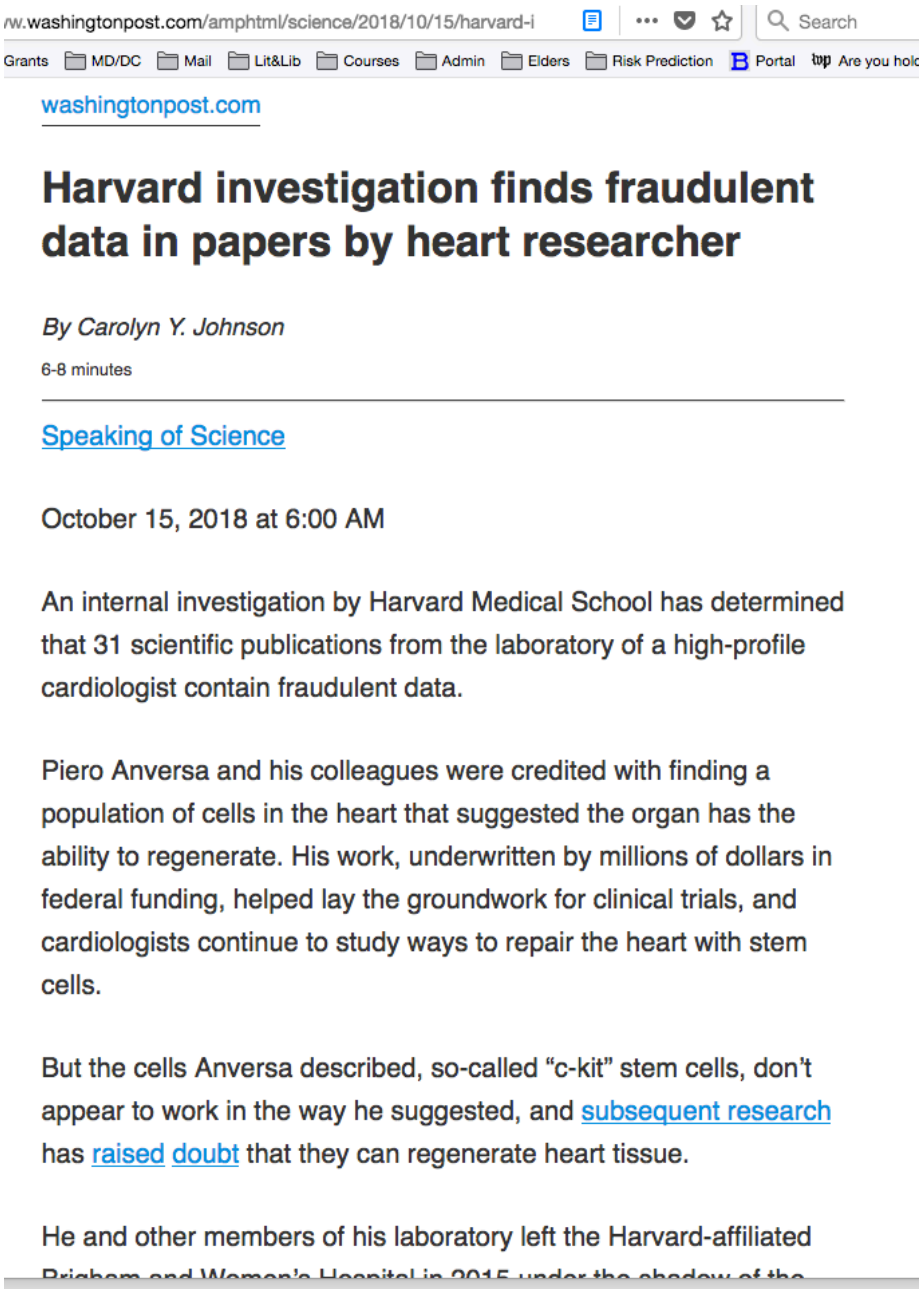
Outline

Objective: What's wrong with data-dredging/ data mining/p-hacking (a rose by any other name)?

- A story (or two) of what could go wrong (among many)
- Different goals in science (etiology vs. prediction)
- What is "it"? Example
- How to properly analyze data (don't forget data cleaning)?
- How to do "it" appropriately?
- Questions and Discussion

A story or Two …

- Very recent
  - https://www.washingtonpost.com/amphtml/science/2018/10/15/harvard-investigation-finds-fraudulent-data-papers-by-heart-researcher/?noredirect=on
  - https://www.nytimes.com/2018/09/29/sunday-review/cornell-food-scientist-wansink-misconduct.html
- Personalized cancer treatment
- Mediterranean diet
- Retraction watch



More Evidence That Nutrition Studies Don't Always Add Up

9-11 minutes

NEWS ANALYSIS

A Cornell food scientist's downfall could reveal a bigger problem in nutrition research.

Not too long ago, Brian Wansink was one of the most respected food researchers in America.

He founded the Food and Brand Lab at Cornell University, where he won attention for studies that showed that small behavioral changes could influence eating patterns. He found that large plates lead people to eat more food because they make portions look smaller and that children eat more vegetables when they have colorful names like "power peas." Dr. Wansink wrote best-selling books and published hundreds of studies. For over a year, he served in a top nutrition policy role at the Department of Agriculture under George W. Bush, where he helped shape the government's influential Dietary Guidelines. His research even led the government to spend almost $20 million redesigning school cafeterias, an initiative known as the Smarter Lunchrooms Movement.

But this month, Dr. Wansink's career at Cornell came to an unceremonious end. On Sept. 20, the university announced that a yearlong investigation had found that he committed "academic misconduct in his research and scholarship, including misreporting of research data," and that he had tendered his resignation. The announcement came one day after the prestigious medical journal JAMA retracted six of Dr. Wansink's studies because of questions about their "scientific validity." Seven of his other papers had previously been retracted for similar reasons.

Broad Downstream Implications

- Results are not reproducible (anything from data quality issue to analysis quality issue)
- Results of meta-analysis not reliable (if only +ve results are published)
- Resources (money and time) wasted
- Science takes a step back

Goals in Research

- Etiologic research: identify association between an agent/exposure and a disease (to be reassessed, reproduced and eventually, to establish cause-effect relationship)
- Hypothesis generation vs. hypothesis testing
- Prediction modeling: predict a disease outcome (and do it well) from however you can (everything and kitchen sink)

Role of Statistics

- Establish association (standard statistical testing, inference, regression)
- Establish causation (causal inference + content knowledge)
- Establish prediction model (and evaluate)

Role of Statistics

- Establish association (standard statistical testing, inference, regression)
- Establish causation (causal inference + content knowledge)
- Establish prediction model (and evaluate)

What is p-value?

- It is NOT the probability that we have truly found an association between exposure/outcome or difference between two (or more) groups
- It is NOT the probability that H0 is true (or H0 is not true)

What is p-value?

- It is the following probability:
  1. Suppose in truth, the groups are the same/similar as reflected by the equality of their means, proportions, etc.
  2. Assuming (1) is true, the probability of getting a result as or more extreme than what we saw in our data is p-value

What is p-value?

- What we would like:

  Pr(No diff in means | data)

- What we get:

  Pr (as or more extreme than our data | No diff in means)

What is p-value?

- Kind of like sensitivity versus predictive value (PV)

Sensitivity = Pr(Test is +ve | Disease)

Positive PV = Pr(Disease | Test is +ve)

Example

- Compare non-motor and motor symptoms between IRBD (D=1) and controls (non-IRBD, D=0)
- UPDRS scores
- Hypothesis: average scores in the two groups are the same (H0: m0 = m1, H1: m0 ≠ m1)
- Significance threshold: 0.01 (alpha)
- Control: 0.2 ± 0.5; IRBD: 1.1 ± 1.5
- p-value= 0.018 (NOT statistically significant)

- Ref: Yao, et al. Longstanding disease-free survival in idiopathic REM sleep behavior disorder: Is neurodegeneration inevitable? Parkinsonism & Related Disorders, 2018.

Example

- p-value= 0.018 – what does it mean?
- If the two groups really had the same mean, the chance of observing the (standardized) mean difference as we saw (in the data at hand) or more "extreme" is 18 in 1000
- "Extreme" in the direction of H1
- H1 could be one or two-sided
- p-value will depend on the one vs. two sided H1

What is “it”?

- p-hacking – search for a statistically significant p-value
- Change alpha
- Change H1 (two-sided to one-sided)
- Test other hypothesis (data dredging: not a priori defined)
  - Same variable, transformed
  - Other variables
  - Being creative
  - etc.

Example

- Compare non-motor and motor symptoms between IRBD (D=1) and controls (non-IRBD, D=0)
- UPDRS II scores
- Hypothesis: average scores in the two groups are the same (H0: m0 = m1, H1: m0 ≠ m1)
- Control: 0.4 ± 0.7
- IRBD: 4.0 ± 3.0
- p-value= 0.002

- Ref: Yao, et al. Longstanding disease-free survival in idiopathic REM sleep behavior disorder: Is neurodegeneration inevitable? Parkinsonism & Related Disorders, 2018.

Example

- Even when there is no difference, the chance of finding a significant result increases with the number of tests performed:

No. tests k	P($\geq$1 Type I error) $1\text{-}0.95^k$
1	5%
2	10%
5	23%
10	40%
15	54%
20	64%
50	92%
100	99%

Broad Downstream Implications

- Results are not reproducible
  - Other labs – repeating same experiment
- Publication bias - results of meta-analysis not reliable
  - Meta-analysis based on only published results
  - Only significant results are published
- Resources (money and time) wasted
- Wheels keep turning
- Science takes a step back

Broad Downstream Implications

Significant p-value does not mean that the results are scientifically meaningful

Questions and Discussion

Best Practice

- Specify, specify, specify
- Analysis plan
  - Based on the protocol
  - Before looking at the data
  - Leave room for exploration, but acknowledge
- Variables and functions/transformations
- Hypotheses: H0, H1 (one vs. two sided)
- Level of significance, confidence interval
- Correct for multiple testing (hint: use a lower threshold)
- Don't test anything that is not listed in the protocol

Best Practice

- Specify, specify, specify
- Analysis plan
  - Based on the protocol
  - Before looking at the data
  - Leave room for exploration, but acknowledge
- Variables and functions/transformations
- Hypotheses: H0, H1 (one vs. two sided)
- Level of significance, confidence interval
- Correct for multiple testing (hint: use a lower threshold)
- Don't test anything that is not listed in the protocol

Best Practice

Together now:

We will NOT test anything that is not listed in the protocol.

Best Practice

But, but, …

… but I still want a significant result!

- Resources invested (time and money)
- PUBLICATION!!!
- How will we generate novel hypotheses if we don't test for "new" things?

… but I still want a significant result!

- Resources invested (time and money)
- PUBLICATION!!!
- How will we generate novel hypotheses if we don't test for "new" things?

What can we do?

- Report properly
- Acknowledge absolutely
  - Start with the protocol (hypothesis testing)
  - State any hypothesis generation activity
- Prediction modeling
  - Test set
  - Validation set (internal/external)

Example: how to do it wrong?

- Get a data: compare “cases” and “controls”
- Don’t have a protocol (or deviate quite a lot from that)
- Try all variables
- Try log transformation
- Try quadratic transformation
- Try a subgroup
- …

Best Practice - Revisited

https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.W0i9Z6lrzdQ

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

Best Practice - Revisited

https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.W0i9Z6lrzdQ

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

Thank you!

- JB, Chun, Elizabeth, Diana
- Funding agencies: NSERC and FRQS

YOU all!

Thank you!

Contact

- Paramita Saha Chaudhuri
- paramita.sahachaudhuri@mcgill.ca
- Google -> paramita biostatistics

https://sites.google.com/site/paramitasaharesearch/

Questions and Discussion