Reproducibility: "Big" Data Analysis and Statistical Methods

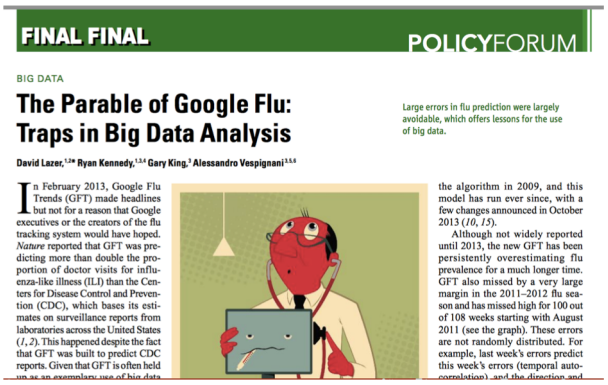
Katerina Kechris, PhD
Professor
Biostatistics and Informatics
Computational Bioscience Program

10/2/2019
BIOS 6606

Outline

- Case Study 1: Flu Trends
- "Re" Definitions – Repeatability, Replicability, Reproducibility
- Case Study 2: Marijuana
- Case Study 3: Vaccines
- Effective Statistical Practice

Case Study 1: Flu Trends



FINAL FINAL

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held

the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal auto-

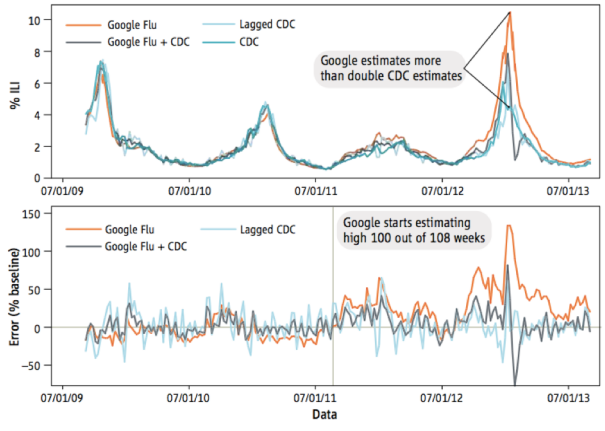
Science, 14 MARCH 2014 VOL 343

Google Flu Trends (GFT) Problems

- Completely missed non-seasonal 2009 influenza A–H1N1 pandemic
- Over-estimated flu prevalence
- Biased by media-panics
- CDC Simple lagged model for flu prevalence better

Google Flu Trends (GFT) Problems

- "Big data hubris"
  - Assumption that big data are a substitute (rather than supplement) to traditional data collection and analysis.
- "Algorithm dynamics"
  - Several changes in Google's search algorithm and user behavior affected GFT's tracking system



Lessons Learned

- Supporting materials not provided for GFT
  - Data, search terms
- Could not be replicated
- "It's not just about size of the data"
- "Transparency and replicability"
- Standards for reproducibility of big data findings

Outline

- Case Study 1: Flu Trends
- "Re" Definitions – Repeatability, Replicability, Reproducibility
- Case Study 2: Marijuana
- Case Study 3: Vaccines
- Effective Statistical Practice

"Re" Definitions

Replication

Repeatability

Reproducibility

http://ivanhanigan.github.io/2016/01/reproducibility-vs-replication-definitional-variations/

Replication

- Independent investigators, methods, equipment, and protocols
- Address the same scientific hypothesis
  - Build evidence *for* or *against* hypothesis
- Essential to scientific method
  - Disciplined approach to scientific discovery
  - Weeds out spurious claims

http://www.replicability.tau.ac.il/index.php/replicability-in-science/replicability-vs-reproducibility.html
http://science.sciencemag.org/content/334/6060/1226

Reasons for Non-Replication

Experimental Issues

- Lack of proper controls (negative or positive)
- Reagents not appropriately validated
- Experiments not repeated
- Insufficient sample sizes
- Results not generalized (e.g. only one cell line or inappropriate model)

Begley & Ellis (2012) *Nature* 483:531-3
Drug development: Raise standards for preclinical cancer research

Reasons for Non-Replication

Analysis & Reporting

- Incorrect/inappropriate statistical analysis
- Incomplete reporting of methods
- Pressure to publish
- Bias towards 'perfect' stories
- Bias against reporting of negative results

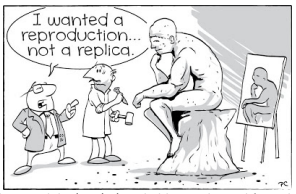
Begley & Ellis (2012) *Nature* 483:531-3
Drug development: Raise standards for preclinical cancer research

"Re" Definitions

Replication

Repeatability

Reproducibility

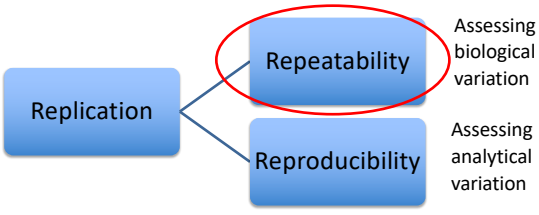


http://ivanhanigan.github.io/2016/01/reproducibility-vs-replication-definitional-variations/

Minimum Standards

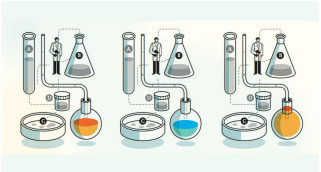
Replication not always feasible (resources, time)

Establish minimum standards for scientific claims:

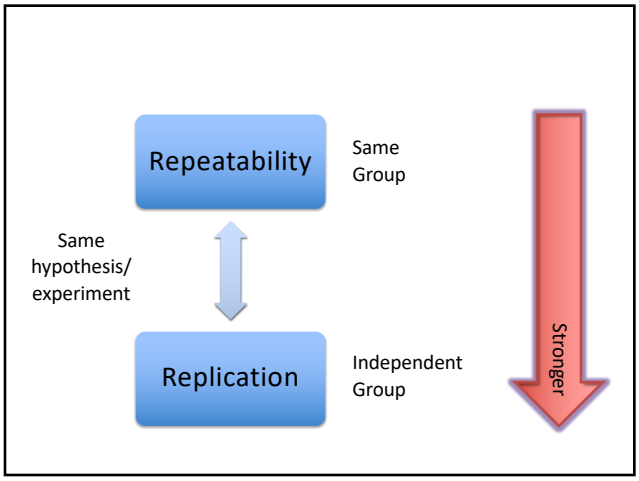


Repeatability

- Agreement between successive measurements carried under same conditions
- Same investigators, methods, equipment, and protocols
- "Closeness"
  - Measured by variance
- Sample size (n>1)



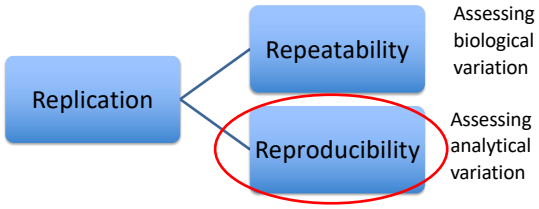
Adapted from NIST Technical Note 1297
https://www.sciencenews.org/article/redoing-scientific-research-best-way-find-truth



Minimum Standards

Replication not always feasible (resources, time)

Establish minimum standards for scientific claims:



Reproducibility

- Computational environment used to derive results are made available
  - data
  - code
  - software
- Independent group, computer, software
- Verify results, figures, tables that are reported

Why?

Promote Best Practices in Science

- Ensure validity of data analysis
- Establish trustworthiness
- Ensure transparency
- Better transfer of knowledge

Sidney Harris (The New Yorker)

Why?

Contributing to Scientific Community

- Sharing of data
- Avoids duplication of effort
- Promotes re-use of methods
- "Inviting" researcher to lab
- Promote new collaborations

https://scientificcommunity.files.wordpress.com/2011/07

Why?

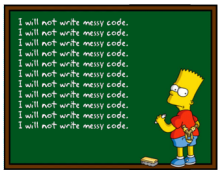
Practical Reasons

- Helps readability and usability – more citations
- Provides a record
- May help find errors or bugs
- Meets journal or grant funding requirements

https://www.notur.no/sites/drupal.uninett.no.notur/

Challenges

- Point & click software - doesn't keep track of analysis
- Code not clean/accessible
  - Too much work to clean up
  - No personnel
  - Not intended for others to see

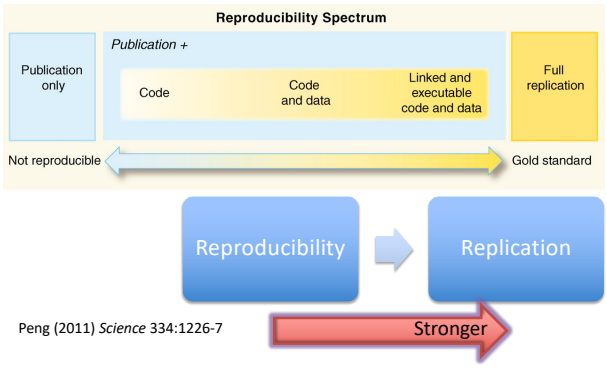


https://assets.firstscribe.com/wp-content/uploads/2015/08/

Challenges

- Didn't work out all the details
  - Only works for special cases
  - Person who knows details, no longer there
- Details & length of manuscript
  - Do readers/referees want to know all the details?
  - Do they have time to check?
- Multiple software packages used
  - alternatives like *snakemake*
- Data not accessible

http://www.ingredientsnetwork.com/47/pdcnewsitem/04/33/52/Patent-61.jpg



Outline

- Case Study 1: Flu Trends
- "Re" Definitions – Repeatability, Replicability, Reproducibility
- Case Study 2: Marijuana
- Case Study 3: Vaccines
- Effective Statistical Practice

Case Study 2: Marijuana Potency

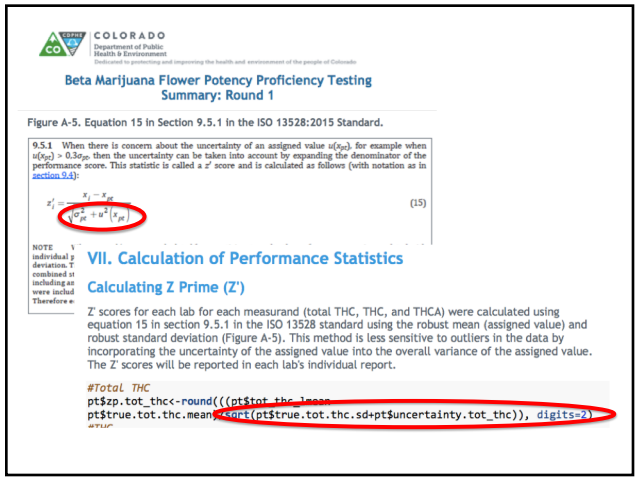
COLORADO
Department of Public Health & Environment

Laboratory Services Division

- Testing and reporting marijuana potency to marijuana retailers
- Accuracy and consistency between laboratories
- No "gold standard" for maximum permissible error to indicate warning and action signals

Purpose

- Invited to be Expert Panelist
- Review 3 rounds of marijuana flower potency testing methods and results across 10 different labs
- Set a maximum permissible error, warning and action signals


COLORADO
Department of Public Health & Environment

Beta Marijuana Flower Potency Proficiency Testing
Summary: Round 1

Figure A-5. Equation 15 in Section 9.5.1 in the ISO 13528:2015 Standard.

9.5.1 When there is concern about the uncertainty of an assigned value $u(x_{pt})$, for example when $u(x_{pt}) > 0.3\sigma_{pt}$, then the uncertainty can be taken into account by expanding the denominator of the performance score. This statistic is called a z' score and is calculated as follows (with notation as in section 9.4):

$$z'_j = \frac{x_j - x_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}} \quad (15)$$

NOTE ... individual ... deviation. T ... combined u ... including an ... were includ ... Therefore e...

VII. Calculation of Performance Statistics

Calculating Z Prime (Z')

Z' scores for each lab for each measurand (total THC, THC, and THCA) were calculated using equation 15 in section 9.5.1 in the ISO 13528 standard using the robust mean (assigned value) and robust standard deviation (Figure A-5). This method is less sensitive to outliers in the data by incorporating the uncertainty of the assigned value into the overall variance of the assigned value. The Z' scores will be reported in each lab's individual report.

```
#Total THC
pt$zp.tot_thc<-round(((pt$tot_thc_lmean-
pt$true.tot.thc.mean)/sqrt(pt$true.tot.thc.sd+pt$uncertainty.tot_thc)), digits=2)
```

Learning lesson

- R markdown report
  - All code documented (no data)
  - Distributed in advance
  - Evaluation & error checking
- Example of Repeatability, Replication & Reproducibility

Outline

- Case Study 1: Flu Trends
- "Re" Definitions – Repeatability, Replicability, Reproducibility
- Case Study 2: Marijuana
- Case Study 3: Vaccines
- Effective Statistical Practice

Case Study 3:

Vaccines

Vaccines

- **Measles Virus:** Can lead to ear infection, pneumonia, seizures, brain damage, and death.
- **Mumps Virus:** Can lead to deafness, meningitis, painful swelling of the testicles/ovaries & rarely sterility.
- **Rubella (German Measles):** During pregnancy could lead to miscarriage or serious birth defects.

- Three-in-one measles, mumps and rubella vaccine (MMR) given since early 1970s in the US
- Nearly eradicated measles & rubella from developed countries

http://www.cdc.gov/vaccinesafety/vaccines/mmr-vaccine.html

Wakefield et al., (1998) *Lancet* 351: 637–641

- Linked MMR with inflammatory bowel disease and "regressive autism", in which language and basic skills were said to have been lost.
  - Author called for boycott of MMR
  - Toured autism conferences
  - "MMR linked with epidemic of autism" (60 minutes)

Wakefield et al., (1998) *Lancet* 351: 637–641

- Caused dramatic decrease in vaccine use (MMR and others)
- Increased incidence of measles and mumps; some resulting in severe injuries & deaths

Public Response

Jenny McCarthy

"In 1983 the shot schedule was ten. That's when autism was one in 10,000. Now there's 36, and autism is one in 150,"

"All arrows point to one direction."

http://briandeer.com/mmr/lancet-summary.htm

Wakefield (1999) *Lancet* 354:949-50

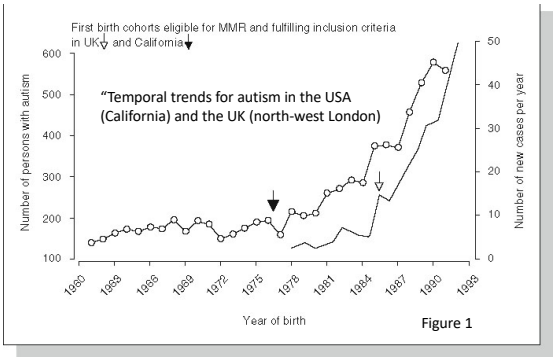
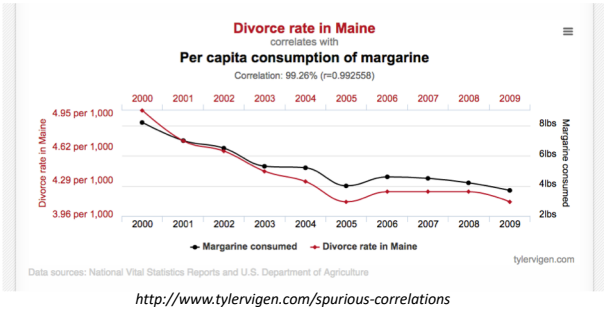


Figure 1

Issues with Graph

Correlation does not cause causation



http://www.tylervigen.com/spurious-correlations

Issues with Graph for CA

- X-axis not autism rates/prevalence for CA
- Original source is Department of Developmental Services (DDS) in CA
  - Enrolled persons with autism
  - Reflects birth rates
  - Used to make case for more funding from legislature – not academic study

Cox & Kirkham (2007) *Drug Safety 30:831-836* "A Case Study of a Graphical Misrepresentation"

More appropriate graph – by age

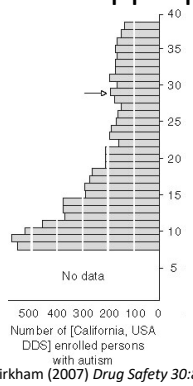


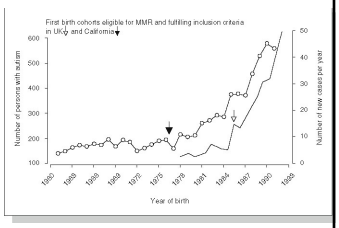
Figure 3:
Arrows indicate the year of the first birth cohort eligible for the measles, mumps and rubella virus vaccine.

In this period, CA population grew nearly by a factor of 2.

Cox & Kirkham (2007) *Drug Safety 30:831-836* "A Case Study of a Graphical Misrepresentation"

Issues with Graph for UK

- Data not comparable with CA
  - UK: Based on medical records, born between 1979-1992, age <=5 before autism diagnosis
  - CA: no exclusion criteria
  - Different scale/starting point



Cox & Kirkham (2007) *Drug Safety 30:831-836* "A Case Study of a Graphical Misrepresentation"

Issues with Graph for UK

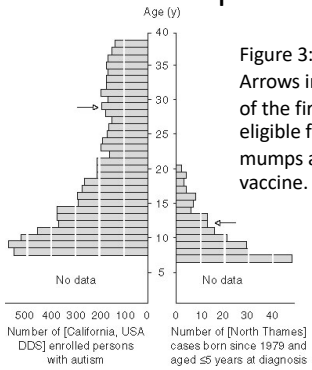


Figure 3:
Arrows indicate the year of the first birth cohort eligible for the measles, mumps and rubella virus vaccine.

Cox & Kirkham (2007) *Drug Safety 30:831-836* "A Case Study of a Graphical Misrepresentation"

Scientifically Unsound & Un-Replicable

- Graphs misleading, original data misrepresented
- Study design inadequate, only n=12!?
- Data incorrect/inaccurate
  - PCR analysis fatally flawed due to contamination; could not possibly detected measles supposed to have detected
  - Colonic histopathology results tainted
  - Clinical diagnosis of the 12 children incorrect

http://briandeer.com/mmr/lancet-summary.htm

Scientifically Unsound & Un-Replicable

- Meta-analyses/reviews performed showed no link between MMR and autism
  - 2004, 2007, 2009, 2012, 2014
  - 100s of studies, millions of children
  - US and abroad (e.g. RAND, NAS, Australia, Italy)

http://briandeer.com/mmr/lancet-summary.htm

2010

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

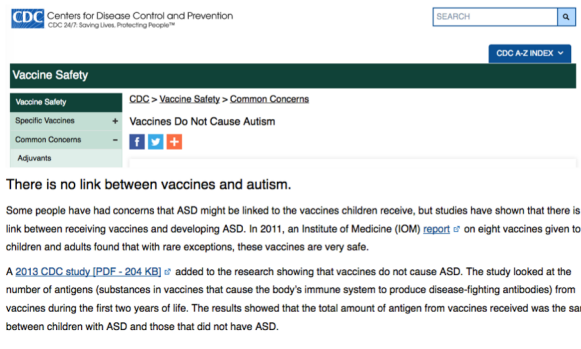
Summary

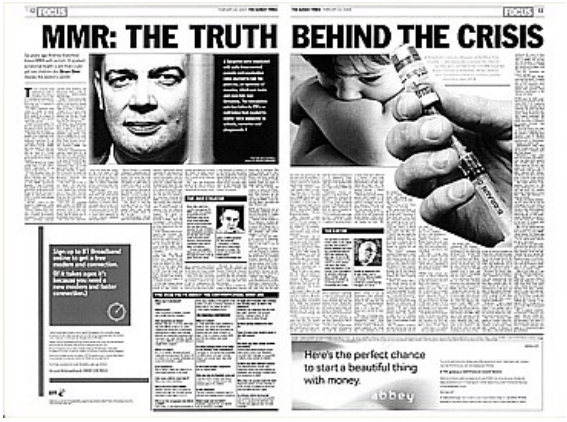
Introduction

Patients and methods

Laboratory investigations

RETRACTED





The Sunday Times (2004)

Fraudulence

- Wakefield had been secretly payrolled to create evidence against the shot
- Planned business schemes meant to profit from the scare (rival vaccine)
- Had concealed, misreported and changed information about the children to rig the results published in the journal

http://briandeer.com/mmr/lancet-summary.htm

UK General Medical Council investigation 2007-2010 lost license: "three dozen charges proved, including four counts of dishonesty and 12 counts involving the abuse of developmentally challenged children"

Anti-vaccine activism – 2015 in CA & OR

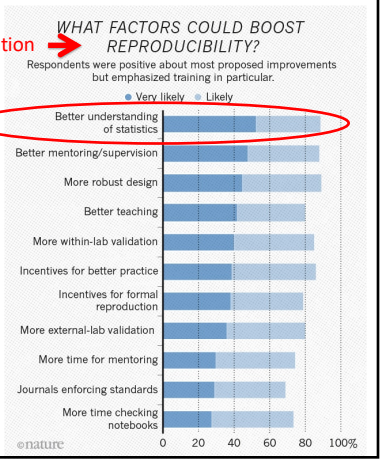
VAXXED – 2016 anti-vaccine propaganda file – withdrawn from Tribeca film festival

Outline

- Case Study 1: Flu Trends
- "Re" Definitions – Repeatability, Replicability, Reproducibility
- Case Study 2: Marijuana
- Case Study 3: Vaccines
- Effective Statistical Practice

Reproducibility & Replication →

Baker (2016) *Nature* 533: 452-4 "1,500 scientists lift the lid on reproducibility: Survey sheds light on the 'crisis' rocking research"



PLOS | COMPUTATIONAL BIOLOGY

(2016) *PLoS Comput Biol* 12(6): e1004961

EDITORIAL

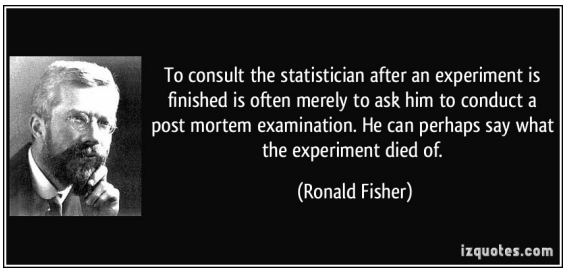
Ten Simple Rules for Effective Statistical Practice

Robert E. Kass[1], Brian S. Caffo[2], Marie Davidian[3], Xiao-Li Meng[4], Bin Yu[5], Nancy Reid[6]*

- Rule 1: Statistical Methods Should Enable Data to Answer Scientific Questions
  - Don't jump to method/test first
- Rule 2: Signals Always Come with Noise
  - Understanding sources of variability

10 Simple Rules

- Rule 3: Plan Ahead, Really Ahead



Kass et al., (2016) *PLoS Comput Biol* 12(6): e1004961

10 Simple Rules

- Rule 4: Worry about Data Quality
  - Data cleaning/munging/carpentry; units/scales; missingness; exploratory data analysis
- Rule 5: Statistical Analysis Is More Than a Set of Computations
  - Rationale & explanation of methods, not just algorithm

Kass et al., (2016) *PLoS Comput Biol* 12(6): e1004961

10 Simple Rules

- Rule 6: Keep it Simple
  - Start simple
  - Add complexity as needed
  - Easier to interpret & explain

Kass et al., (2016) *PLoS Comput Biol* 12(6): e1004961

10 Simple Rules

- Rule 7: Provide Assessments of Variability
  - Report statistical uncertainty (standard errors, confidence intervals)
  - Need to account for extra sources of variability (batches, technical effects, etc)
  - Dependencies in the data

Kass et al., (2016) *PLoS Comput Biol* 12(6): e1004961

10 Simple Rules

- Rule 8: Check Your Assumptions
  - Distribution? Linear? Biases in data collection? Confounding? Check how will model fits the data (diagnostics & graphing)
- Rule 9: When Possible, Replicate!
  - If not, data perturbations
  - Otherwise, at the minimum ….
- Rule 10: Make Your Analysis Reproducible

Kass et al., (2016) *PLoS Comput Biol* 12(6): e1004961

http://www.jir.com/