BALSAM

Finding Significance

Insights from the BALSAM Network

Jeff Bakal, PhD, PStat & Cindy Westerhout, PhD
CRSS
June 13, 2019

BALSAM

Jeff Bakal, PhD, PStat

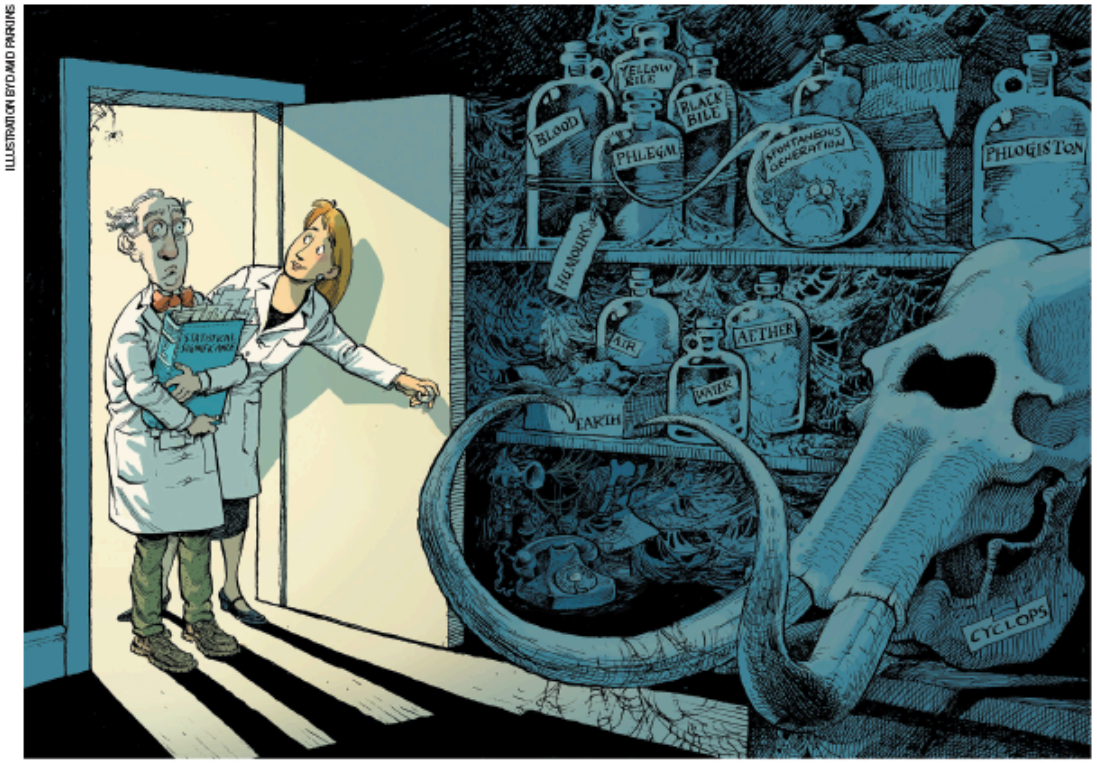
- AHS Director - Provincial Research Data Services
- (Self) proclaimed Science Computerist

Cindy Westerhout, PhD

- Associate Director (Research & Strategic Planning) at the Canadian VIGOUR Centre
- Epidemiologist

BALSAM

A(nother) Call to Retire Statistical Significance



Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

- Not the first
- But has generated discussion, esp on SoMe
- +800 signatories

...Activist-statisticians

Nature 2019

BALSAM

LAW360

Portfolio Media. Inc. | 111 West 19th Street, 5th floor | New York, NY 10011 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

New Views On Statistical Significance Affect Expert Testimony

By **Josh Becker, Aaron Block and Patrick Hill** (May 23, 2019, 10:37 AM EDT)

Statistical significance, a concept often invoked to help characterize the strength or weakness of scientific results, is a prominent feature in modern litigation. Cases requiring expert testimony on the issue of causation, for example, often involve analyses of statistical significance to support or attack (and potentially exclude) an expert's causation conclusion.

What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?

John P. A. Ioannidis

To cite this article: John P. A. Ioannidis (2019) What Have We (N of Scientific Papers with *P* Values?, The American Statistician, 73
10.1080/00031305.2018.1447512
To link to this article: https://doi.org/10.1080/00031305.2019.1

THE AMERICAN STATISTICIAN

IN FOCUS NEWS

REPRODUCIBILITY

THIS MONTH

Statisticians issue warning on *P* values

Statement aims to halt missteps in the quest for certainty.

BY MONYA BAKER

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

cannot indicate the importance of a finding; for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.

Giovanni Parmigiani, a biostatistician at the Dana Farber Cancer Institute in Boston, Massachusetts, says that misunderstandings about what information a *P* value provides often crop up in textbooks and practice manuals. A course correction is long overdue, he adds. "Surely if this happened twenty years ago, biomedical research could be in a better place now."

POINTS OF SIGNIFICANCE

P values and the search for significance

Little *P* value
What are you trying to say
Of significance?

—Steve Ziliak

The significance of experimental results is often assessed using *P* values and estimates of effect size. However, the interpretation of these assessment tools can be invalidated by selection bias when testing multiple hypotheses, fitting multiple models or even informally selecting results that seem interesting after observing the data. Our goal this month will be to identify some circumstances that can give rise to such questionable practices—broadly termed '*P* value hacking' and 'data dredging'. In addition, statistically significant results may not translate into biologically meaningful conclusions—with large sample sizes or small variability, even tiny effects can be statistically significant.

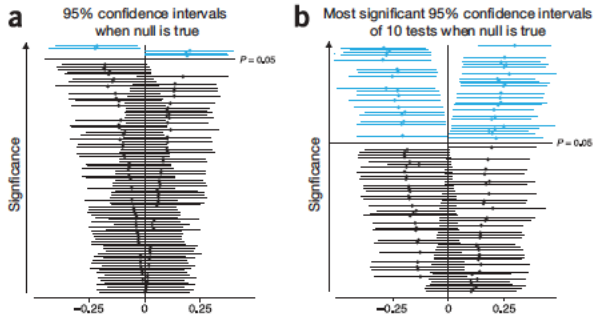


Figure 2 | Merely reporting 95% confidence intervals does not address selection bias. (**a**) 95% confidence intervals for 100 one-sample *t*-tests with samples of size $n = 100$, mean zero and s.d. = 1. Intervals are vertically sorted in increasing order of statistical significance. (**b**) 100 instances of the 95% confidence interval corresponding to the most significant result from a set of 10 one-sample *t*-tests of the kind performed in **a**.

experiment. In reporting the most significant *P* value, we are actually considering the distribution of the minimum of 10 random uniform distributions (**Fig. 1b**). This distribution is readily computed and has density $k(1-x)^{k-1}$ for k independent tests. Using $k = 10$, the probability of observing $P < 0.05$ is $1 - (1 - 0.05)^{10} = 0.40$ (**Fig. 1b**).

Reporting a statistically significant result as if this were the only test performed is an example of selection bias and leads to inflated

BALSAM

The Vicious Cycle

Q: Why do so many colleges and grad schools teach p=0.05?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use p=0.05?

A: Because that's what they were taught in college or grad school.

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

George Cobb, 2014

ASA p-value Statement

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

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.*

BALSAM

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

ASA *P*-Value Statement Viewed > 150,000 Times

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'"

"Over time it appears the *p*-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect,' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as '*p*-hacking' and 'data dredging' that emphasize the search for small *p*-values over other statistical and scientific reasoning."

BALSAM

Scope of the "Problem"

- 51.1% of PubMedCentral abstracts (1990-2015) had p-values for NHST in abstract or text
  - Random sample (n=100): 4 report CIs on effect sizes, 0 Bayesian, 0 FDR methods
- Higher rate in clinical journals and in meta-analyses

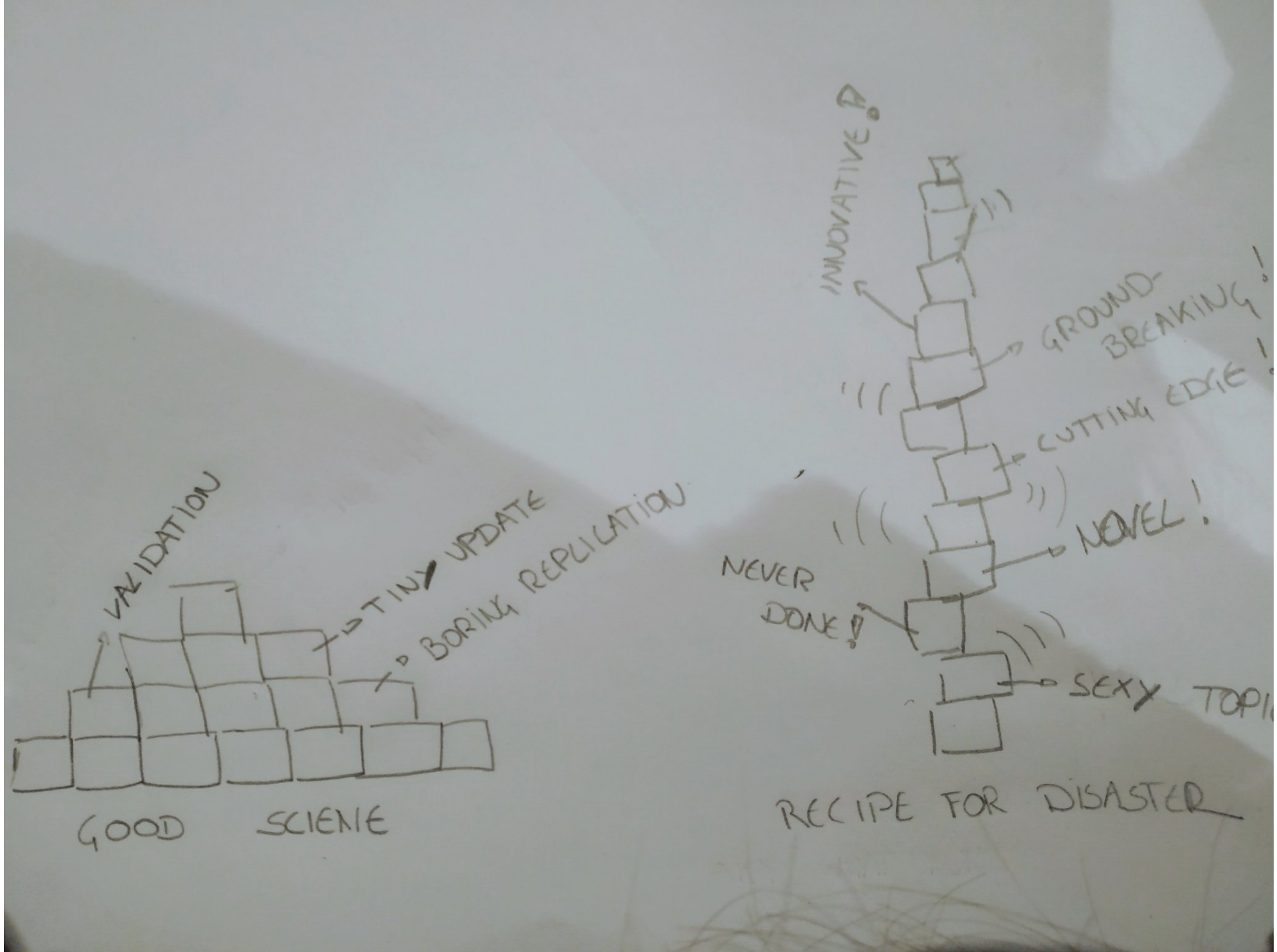
Overall (all papers)	51.1%
Articles	78.4%
Meta-an	82.8%
Random	76.0%
Other clinical trials (exc	

So what? Selection bias!

Which p-values are reported and where….

Ioannidis 2019

BALSAM


VALIDATION
TINY UPDATE
BORING REPLICATION
GOOD SCIENCE
INNOVATIVE!
GROUND-BREAKING!
CUTTING EDGE!
NOVEL!
NEVER DONE!
SEXY TOPI
RECIPE FOR DISASTER

BALSAM

Early days of the p-value

- RA Fisher introduced it as a formal research tool but without defining inferential meaning
  - A rough numeric guide of the strength of evidence against the null hypothesis
  - An evidential tool to be used flexibly within the context of a given problem.
  - Proposed the use of "significant" to be attached to small p-values

'Personally, the writer prefers to set a low standard of significance at the 5 percent point…a scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.'

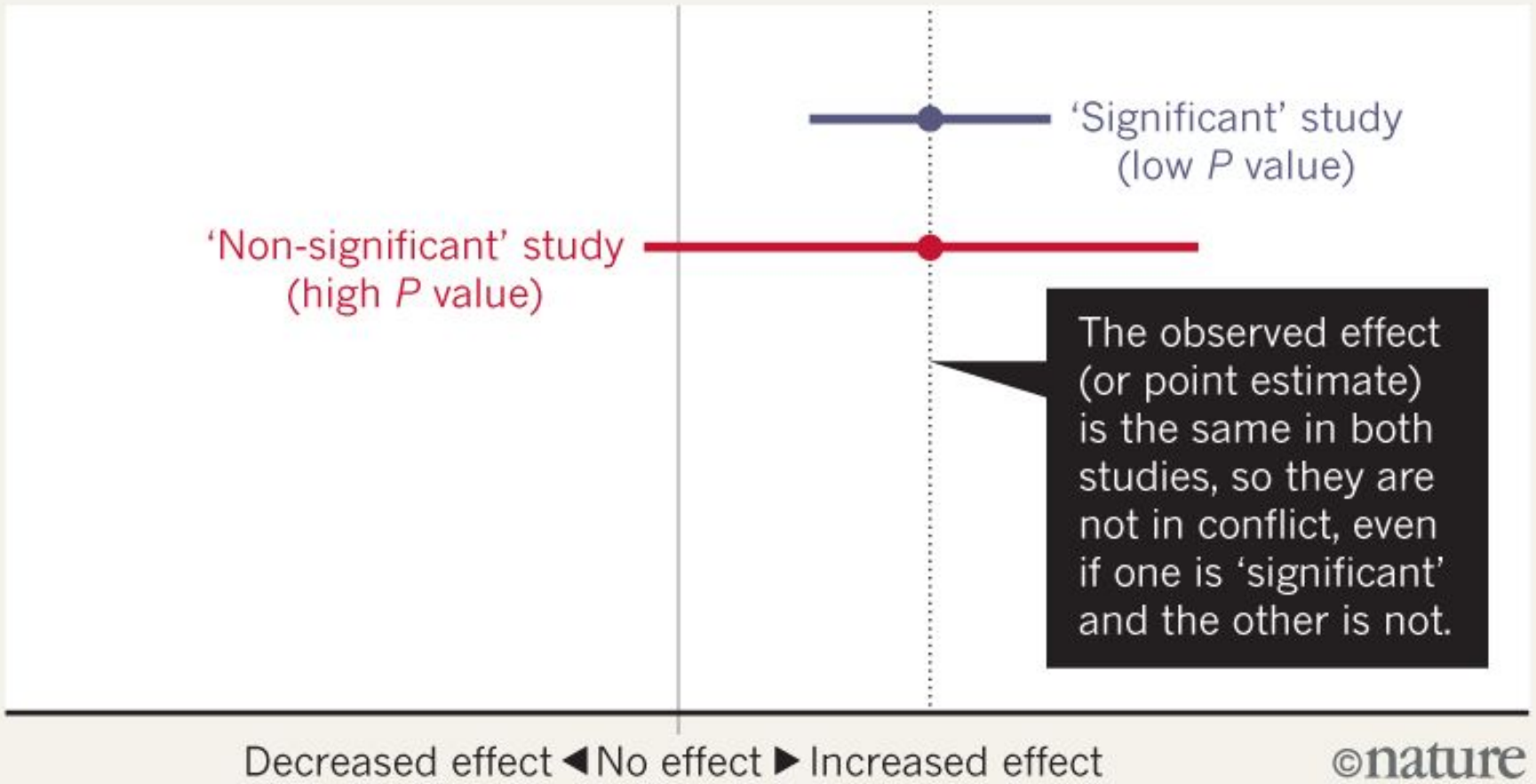
- Operationally, p-value<0.05: Repeat the experiment! Not proof itself!

Goodman Semin Hematol 2008

BALSAM

BEWARE FALSE CONCLUSIONS

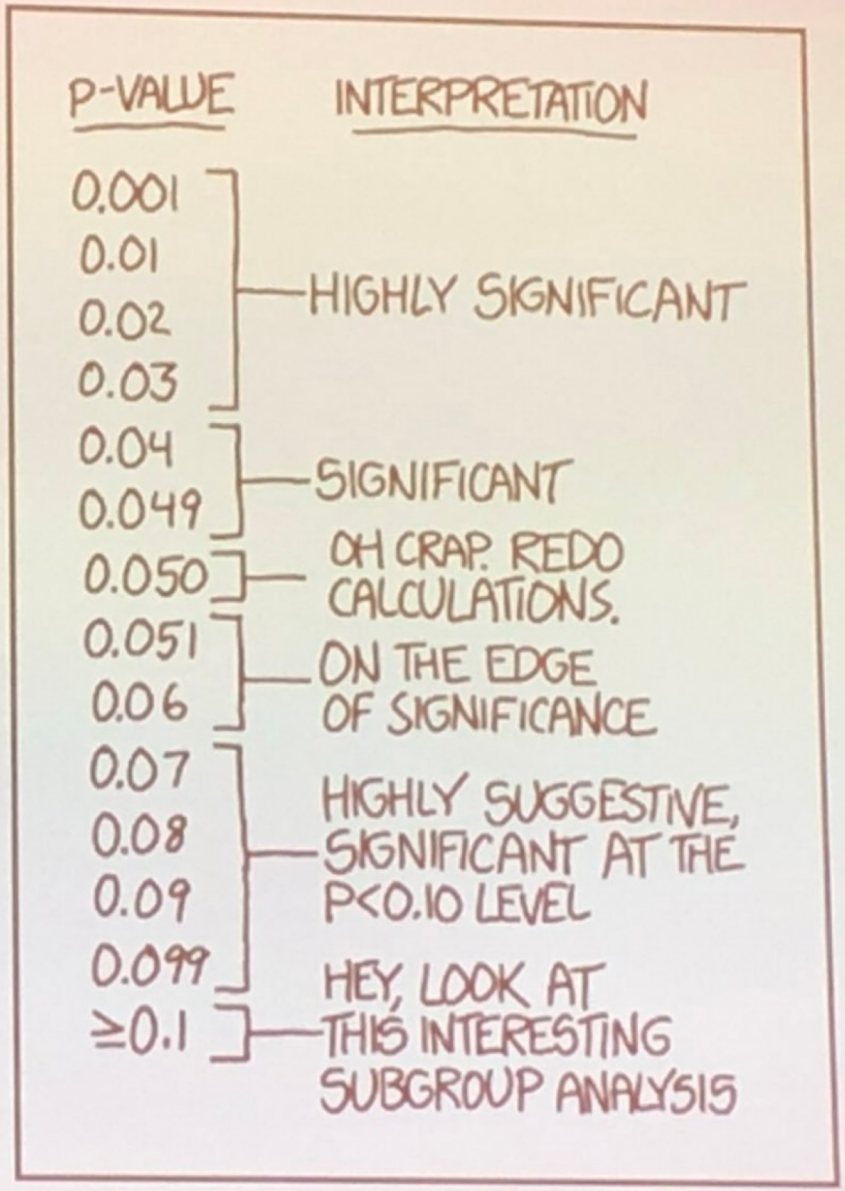
Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.



Amrhein, Greenland, McShane 2019

BALSAM

The problem with p=0.05 or how **not** to interpret p values



Twitter @Thomas_Hiemstra

BALSAM

And today…

Ellie Murray liked

🔥Kareem Carr🔥 @kareem_carr ·

Top 3 p-values ranked:

3. Exactly 0.05 A mystical experienc
edge

2. 10^-62 Unexpected and delightfully optimistic. A thrill.

1. 0.0499 Like a brush with death. It makes you FEEL ALIVE!!!

#epitwitter #statstwitter

5 10 79

BALSAM

Our frienemy the p-value

- Probability of obtaining the value you did or more extreme given the null hypothesis.
- Used extensively
  - “We teach it because it’s what we do; we do it because it’s what we teach.”
- Historically to determine Joy, publication, tenure track, grant renewal (<0.05)
- Despair, ruin, search for additional controls (>0.05) (RR 1989)

BALSAM

It's not the p-value… it's the people

- Humans like to dichotomize things.

- If someone has BP > 140/90 -> Hypertension
  - Is someone at 139/90 clinically significantly different than 141/90?

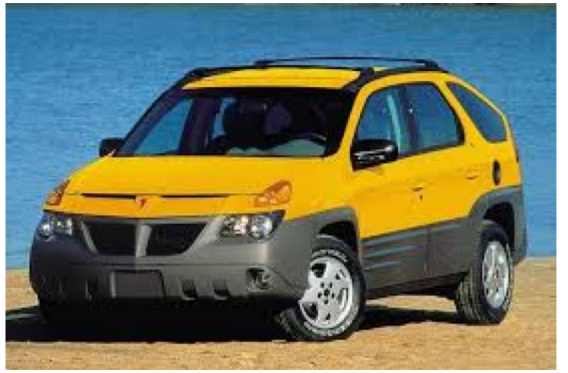
- A need to answer does Rx work?
  - Need to set a line.
  - Regulatory

"That is we want to underscore that, surely God loves the 0.06 nearly as much as the 0.05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p? "–Rosnow and Rosenthal 1989

BALSAM

Prediction Feature Optimization

- People also prefer A caused B
- Important to understand that some things aren't deterministic
- Humans like to believe they can manufacture:
  - Olympic Medals (without cheating)
  - Top 40 records
  - Viral Videos
  - Best selling cars
  - Patient outcomes



Offroad! Sporty! Fun!

BALSAM

Proposed Solutions

- Some suggest changing the limit

- Multiple test corrections
  - Bonferroni, etc
- Alternate Methods

- If behavioral economics taught us anything, it is that in most reactions and corrections, there is often a tendency for over-reaction and over-correction.

BALSAM

Evolving areas

- Renewed interest in using “non statistical” AI techniques
  - (developed mostly by statisticians)

- Very good at classification, without a model prediction becomes hard, as there is no foundation.
  - Classifying the dead doesn’t require “Big Data”
  - Predicting is hard - certain/time accurate prediction is impossible.

BALSAM

Replication "Crisis"

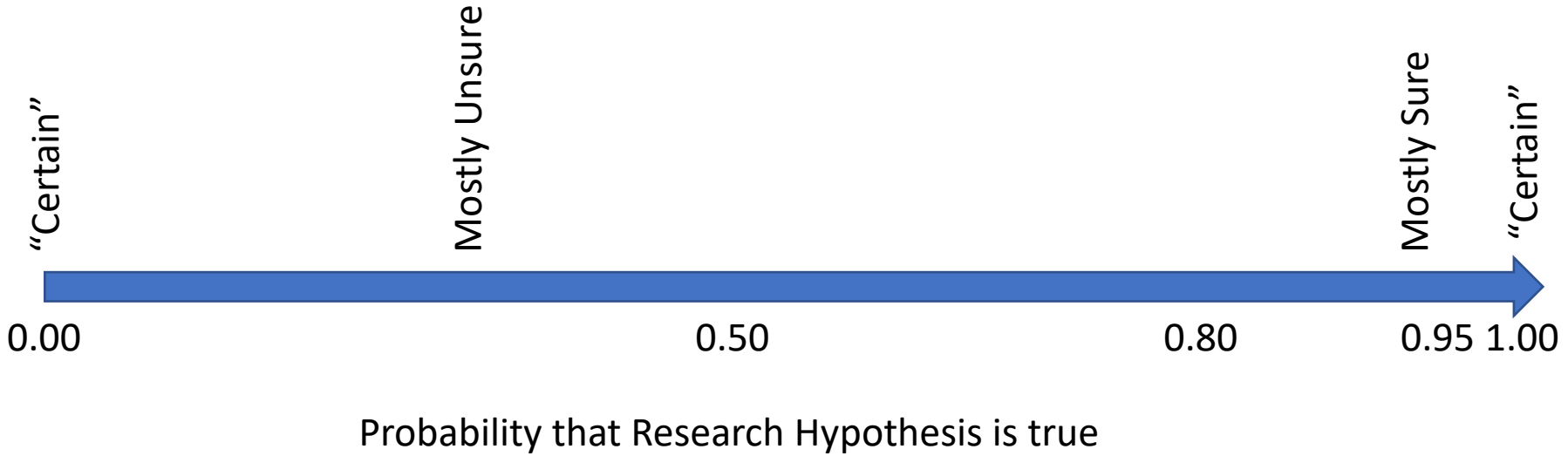
- "Reward systems exist that are fouled up in behaviours which are rewarded are those which the rewarder is trying to discourage"

  -On the Folly of Rewarding A, while hoping for B – Kerr (1995)

BALSAM

If you meet Buddha on the Road, kill him (her)

- There are many alternate religions around this, end of the day still about the March of Science.
- Move from imprecise “mostly unsure” to precise “mostly sure”.



Ralph O’Brien

Signal to Noise

- Important part is that ultimately we are separating trying to pick out the signal from the noise
- Not yes/no
- Need clinical relevance
- Good measurement
- Underlying model

BALSAM

WHAT MAKES ANALYTICS MEANINGFUL?

- DEEPLY DEFINE THE CHALLENGE AT HAND.
- EXAMINE ALL AVAILABLE DATA SOURCES.
- LEVERAGE TOOLS AND ALGORITHMS THAT ADDRESS THE QUESTION.
- ARRIVE AT AN IMPACTFUL DECISION.

BALSAM

How do we proceed?

- Walk through a Study

BALSAM

Idea Generation, Planning and Implementation

- PI develops idea and background

- Works with biostatisticians to develop a grant
- Identify Methods and Data Sources
- Data Validation

BALSAM

Research Products

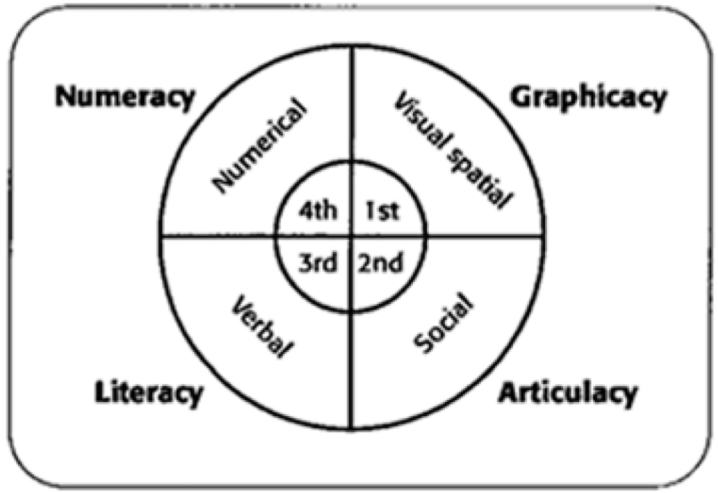


Figure 2. Balchin's "four types of ability."

numeracy —formulating and solving problems using mathematics and computing

articulacy —speaking and listening; also people skills

literacy —writing and reading

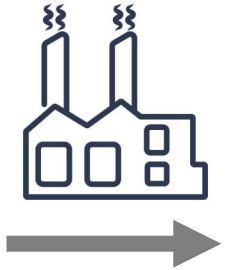
graphicacy —producing and understanding graphics

W.G.V. Balchin 1976

BALSAM

Manuscript Results – Table 1

- Table 1
- Do we need p-values?
  - Are you going to do anything about it?
  - Are you really testing hypotheses?
- Large studies (everything is significant)

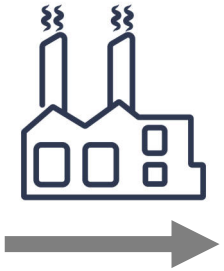


Help lay out summary tables and figures

BALSAM

Manuscript Results Table 2,3,4…

- Identify appropriate analyses accurately answer the research question
- Integration of the results into the clinical realm
- Meaningful interpretations of the Results
- Visualizations that tell the story



BALSAM

Interpretations…

H_0: Mean SBP_B – Mean SBP_A = 0

	Treatment A	Treatment B
		
Mean SBP:	**134mmHg**	**130mmHg**

Accurate Statement That is More Direct

- Assuming the study's experimental design and sampling scheme, the probability is 0.4 that another study would yield a test statistic for comparing two means that is more impressive that what we observed in our study, if treatment B had exactly the same true mean SBP as treatment A. (This statement is best at pointing out the limitations of how p-values can be interpreted.)

Shorter Statement

- The study did not contradict the supposition that treatments A and B yield the same mean SBP (p=0.4).

Harrell, 2018

BALSAM

Absence of Evidence ≠ Evidence of Absence

- **DO NOT** use arbitrary thresholds for significance and to not attempt to use the p-value by itself to present evidence from data about underlying effects.

- **DO** include ***compatibility interval*** in the text.
  - The set of all values of the true effect that are compatible with the data in the sense of not rejecting a hypothesis.

Harrell, 2018

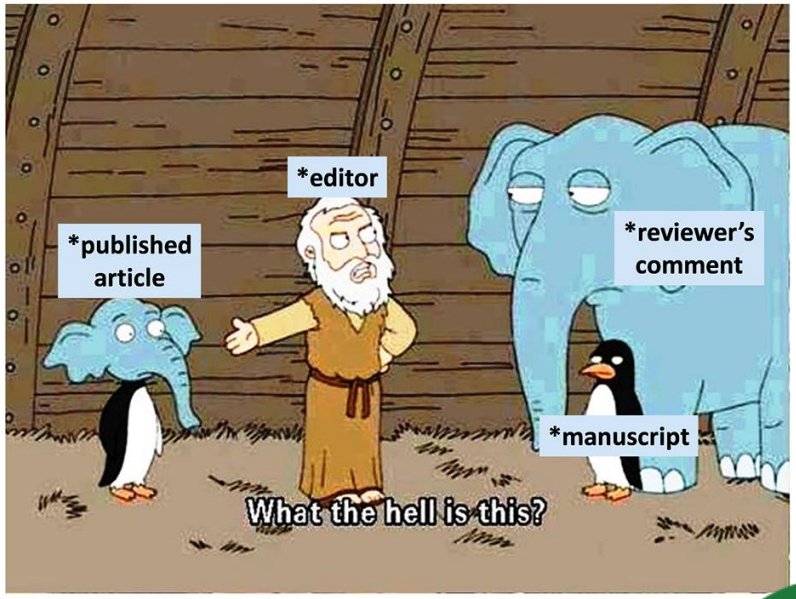
BALSAM

After the Review(s)

- Work with reviewer comments

- Use sound methodology to address issues
- Ensure that the ‘plot’ isn’t lost

Try to prevent



BALSAM

What to do…

A **A**ccept uncertainty

T be **T**houghtful

O **O**pen

M **M**odest

‘Statistically Significant’
Don’t Say It and Don’t Use It!!

Wasserstein, Schirm, Lazar 2019

BALSAM

What to do…

- Enhance Methods sections and data tabulation with more detail and nuance.
- Emphasize estimates and uncertainty in them.
- Press for results interpretation and publication that are no longer based on statistical thresholds.
- Spend less time with statistical software and more time thinking.

Amrhein, Greenland, McShane 2019

BALSAM

Who is out there to help?

Accessible resources…

BALSAM

STRATOS INITIATIVE

STRengthening Analytical Thinking for Observational Studies

- *An efficient way to help researchers to keep up with recent methodological developments is to develop guidance documents that are spread to the research community at large.*
- A large collaboration of experts in many different areas of biostatistical research.
- *STRATOS: To provide accessible and accurate guidance in the design and analysis of observational studies.*

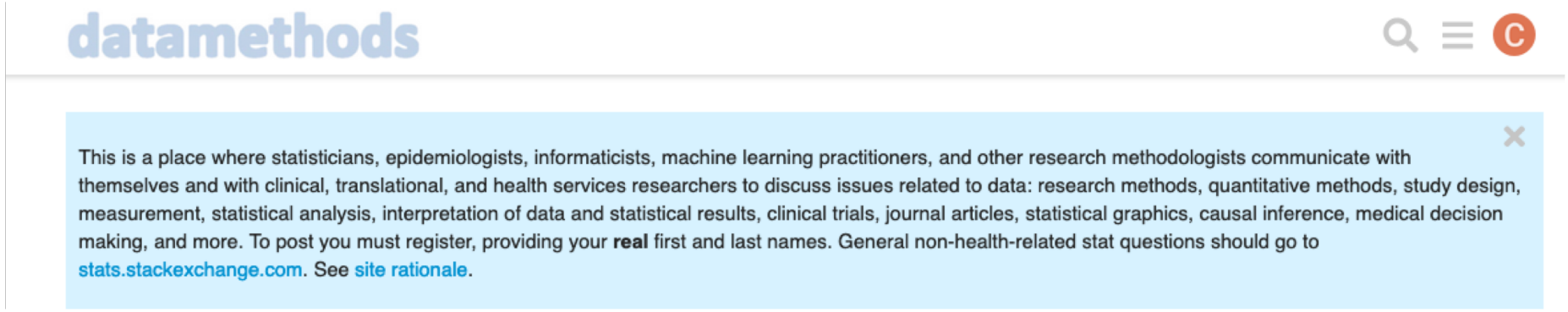
Topic groups

1	Missing data
2	Selection of variables and functional forms in multivariable analysis
3	Initial data analysis
4	Measurement error and misclassification
5	Study design
6	Evaluating diagnostic tests and prediction models
7	Causal inference
8	Survival analysis
9	High-dimensional data

BALSAM

Online…

- ASA AMERICAN STATISTICAL ASSOCIATION Promoting the Practice and Profession of Statistics®
  Société statistique du Canada Statistical Society of Canada
- #biostat, #epistat
- **datamethods.org** Forum: Online chatroom with many active practitioners of statistics, including well-recognized experts



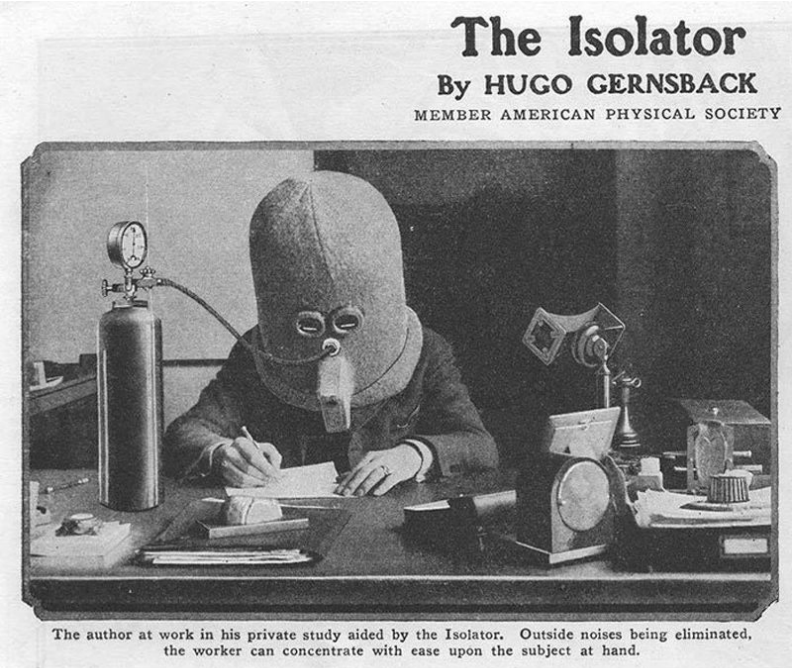
BALSAM

Biomedical **A**na**L**ytics
Studio **a**nd **M**ethods Network

BALSAM

Why BALSAM?

- Applied biostatisticians and data analysts are isolated and missing an Academic Home
  - ⛔ sharing knowledge, developing resources, & academic pursuits
  - Jack-of-all tracks vs masters
  - Professional development & advancement is challenging



- Access to skilled analytic collaborators is imbalanced and difficult

BALSAM

Who is BALSAM?

- Nearly 30 practitioners of statistics in clinical research within FoMD
  - Data/statistical/research analysts, (bio)statisticians, mathematicians, epidemiologists, health economists
- Mix of early career to senior level
  - Masters, PhDs
- Single-stat settings (N=1) to stats teams (N=10)

Alberta SPOR SUPPORT Unit

Canadian VIGOUR Centre
Bridging Hearts and Minds

Kidney Health Research Group

Alberta Health Services

PEDIATRICS

NACTRC Real World Evidence

women & children's health research institute

Cardiac Surgery

BALSAM

Areas of Expertise

BALSAM

Data Expertise

CLINICAL TRIALS

PROVINCIAL ADMIN

HEALTH DATA

NATIONAL HEALTH DATA

REGISTRIES

BALSAM

Core Values

Collaboration

To create a platform for *partnership* with clinical investigators and advancing clinical research.
To establish an academic home for methodologic research led by biostatisticians and data scientists.

Excellence

To raise the level of knowledge and practice related to the use of data science, study design, statistical methods and analysis in clinical research.

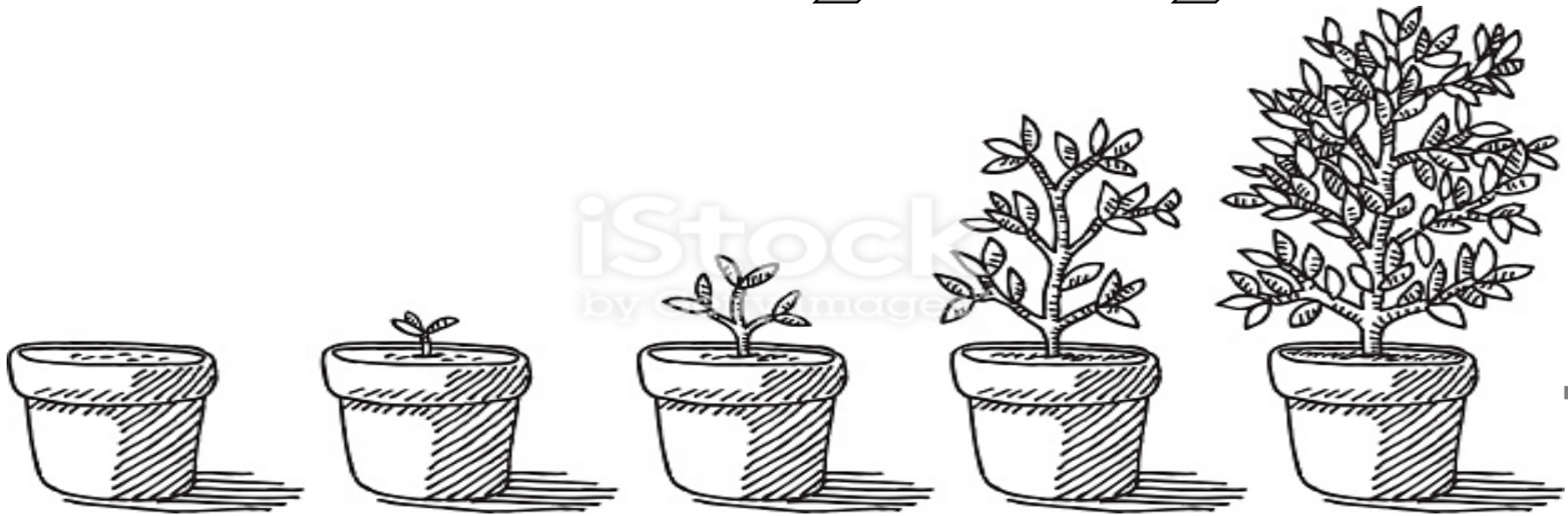
Efficiency & Expertise

To promote the development of and specialization in analytic tools through a coordinated team approach.

BALSAM

Every step of the way…

CONCEPT › DESIGN › DATA › ANALYSIS › INTERPRETATION › SHARING



ARCHIVAL

TRACKING

BALSAM

What must BALSAM do…

Short Term

- Provide a hive for the analytics community
  - Enable collaboration
  - Home for analytic innovation

- Attract and retain high-quality members
  - Community and PD

Longer term

- Support the *Research Studio*
  - develop and refine clinical investigation
- Pair clinical and analytic teams
  - Use best available analytics techniques to generate sound answers to health research questions

BALSAM

What BALSAM could do…

- A ‘record label’ for Data Science in the FoMD
  - Added credibility
  - ‘Stamp of methodological approval’ for grants, papers and study design
- Link between AHS and UA as part of a *learning healthcare system*
- Become a partner in advancement of clinical research at the UA

What it isn’t…

…a general research platform.

BALSAM is for data analysts, biostatisticians and data scientists at the UA.

BALSAM

Now and What lies ahead…

- ~30 members connected
- Established a monthly forum for discussing practical problems/solutions and a journal club

- Slack channel to keep members informed

- Arranging/facilitating professional development and training
- A members' registry of expertise and interests is *in progress*

Fin

“The most successful organizations create an environment that is hospitable to risk-taking, innovation, and creativity.” - Donald Rumsfeld

BALSAM

Thank you!