Moving to a World Beyond p<0.05

Ron Wasserstein

Executive Director

American Statistical Association

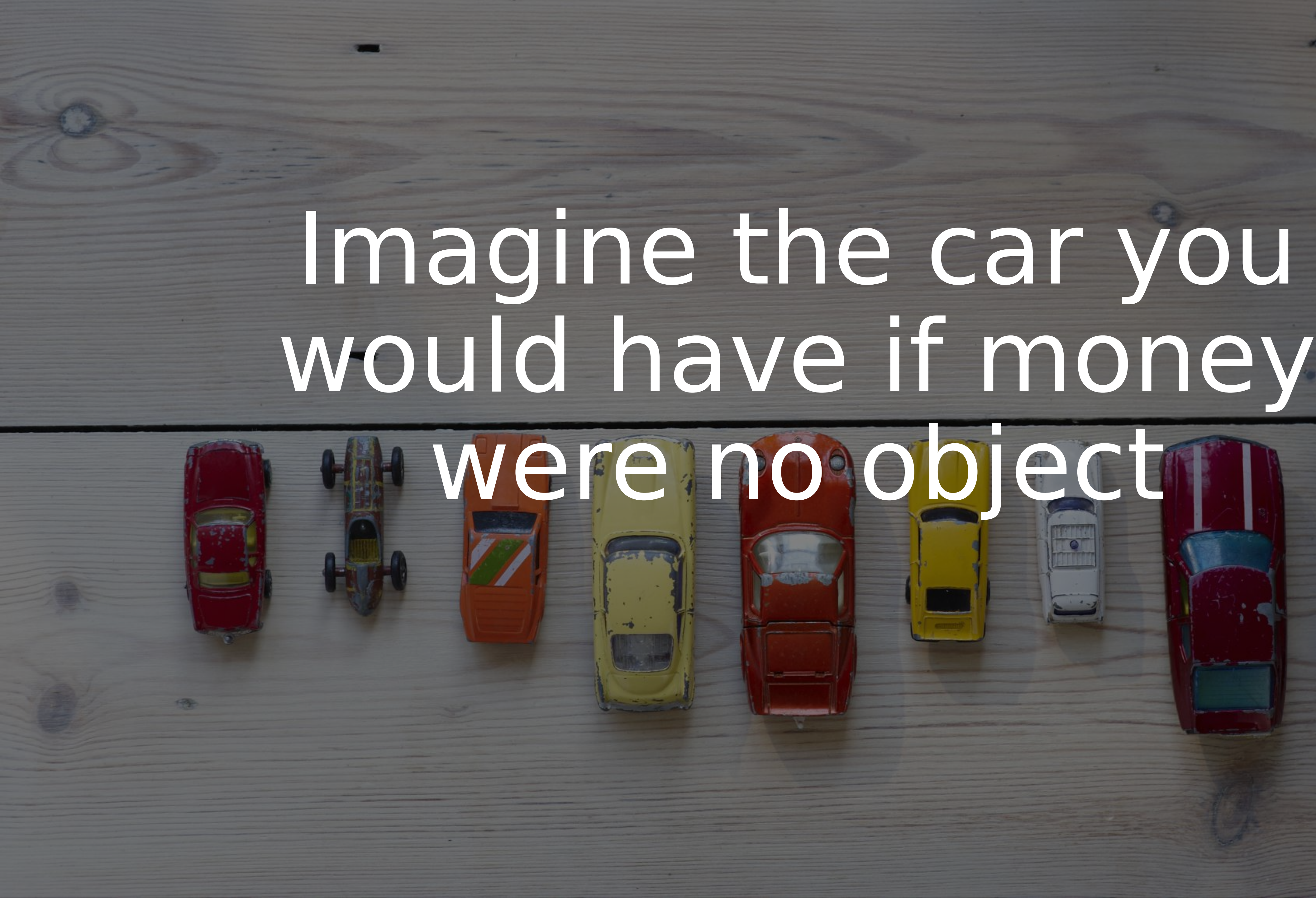
May 31, 2024

Milwaukee Chapter Annual Meeting

DISCLAIMER:

I am here today speaking
individual researcher and
capacity as Executive Dire

So, blame me, and not the
anything I say that you do



Imagine the car you would have if money were no object

RR

S MB 7252

“Shaken, not stirred”

“Sometimes it's

"He stole John Wick’s car, sir. And, uhhh, killed his dog."

“Just remember this – in this country they drive on the wrong side of the rode.”

“W

RENAULT
PETRONAS
Red Bull
Red Bull
TOTAL
TOTAL
PHILIPS

“Who you gonna call?

“Ka-Chow!”

"What about the accent? Is it… is it too much?"

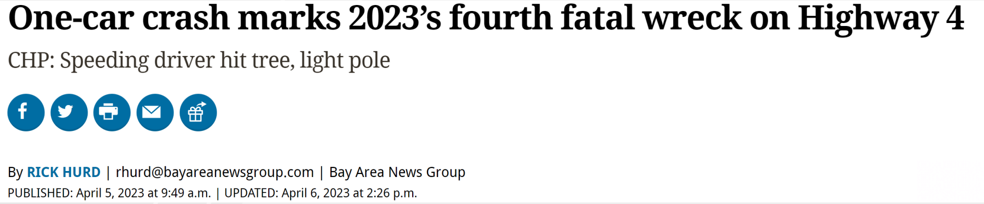
“I am Iron Ma

"It's not who I
underneath, b
do that define

“Are you telling me you built a time machine...out of a De

Suppose had the r amazing ever…

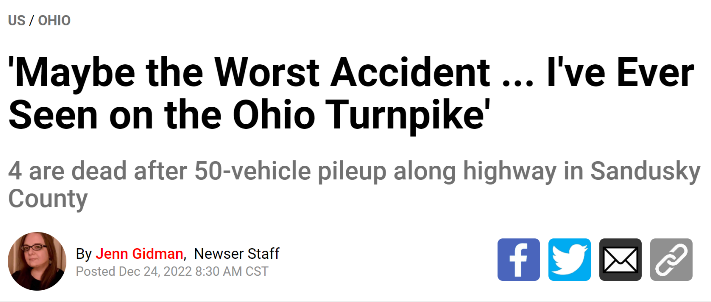
- Beautiful
- Energy efficient
- Everyone has acc
- But…it turns out to drive



One-car crash marks 2023's fourth fatal wreck on Highway 4

CHP: Speeding driver hit tree, light pole

By RICK HURD | rhurd@bayareanewsgroup.com | Bay Area News Group
PUBLISHED: April 5, 2023 at 9:49 a.m. | UPDATED: April 6, 2023 at 2:26 p.m.

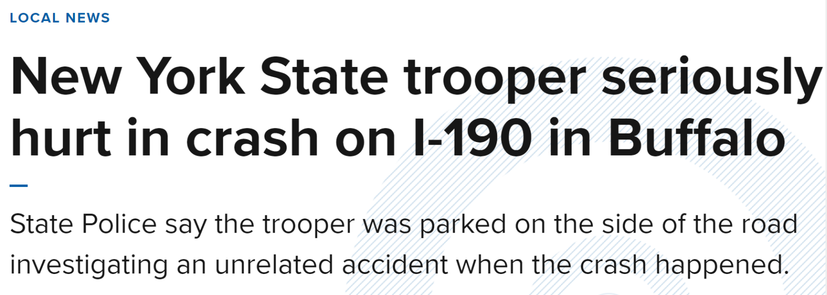


US / OHIO

'Maybe the Worst Accident ... I've Ever Seen on the Ohio Turnpike'

4 are dead after 50-vehicle pileup along highway in Sandusky County

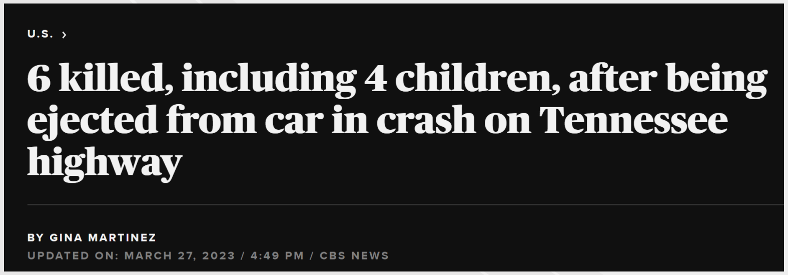
By Jenn Gidman, Newser Staff
Posted Dec 24, 2022 8:30 AM CST



LOCAL NEWS

New York State trooper seriously hurt in crash on I-190 in Buffalo

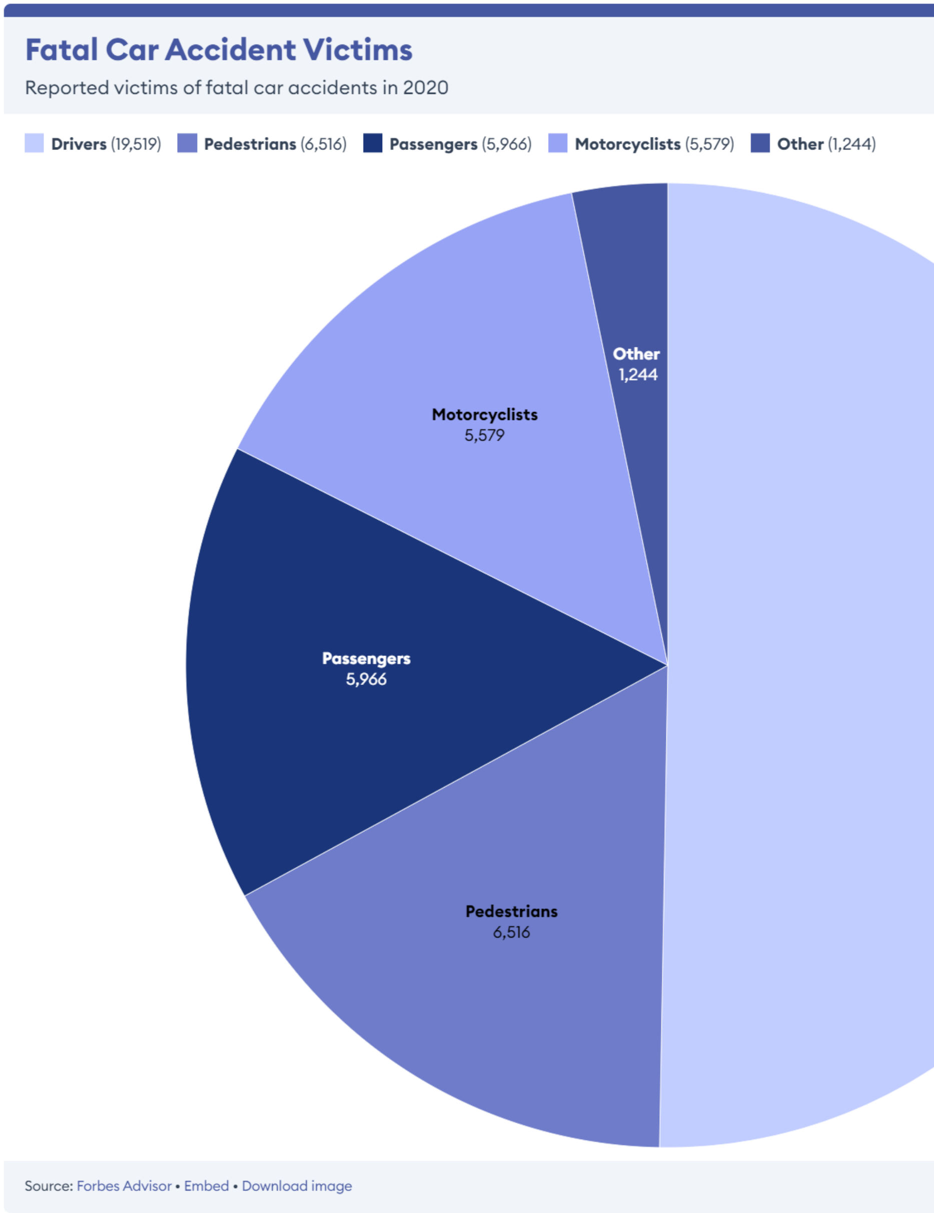
State Police say the trooper was parked on the side of the road investigating an unrelated accident when the crash happened.



U.S. ›

6 killed, including 4 children, after being ejected from car in crash on Tennessee highway

BY GINA MARTINEZ
UPDATED ON: MARCH 27, 2023 / 4:49 PM / CBS NEWS

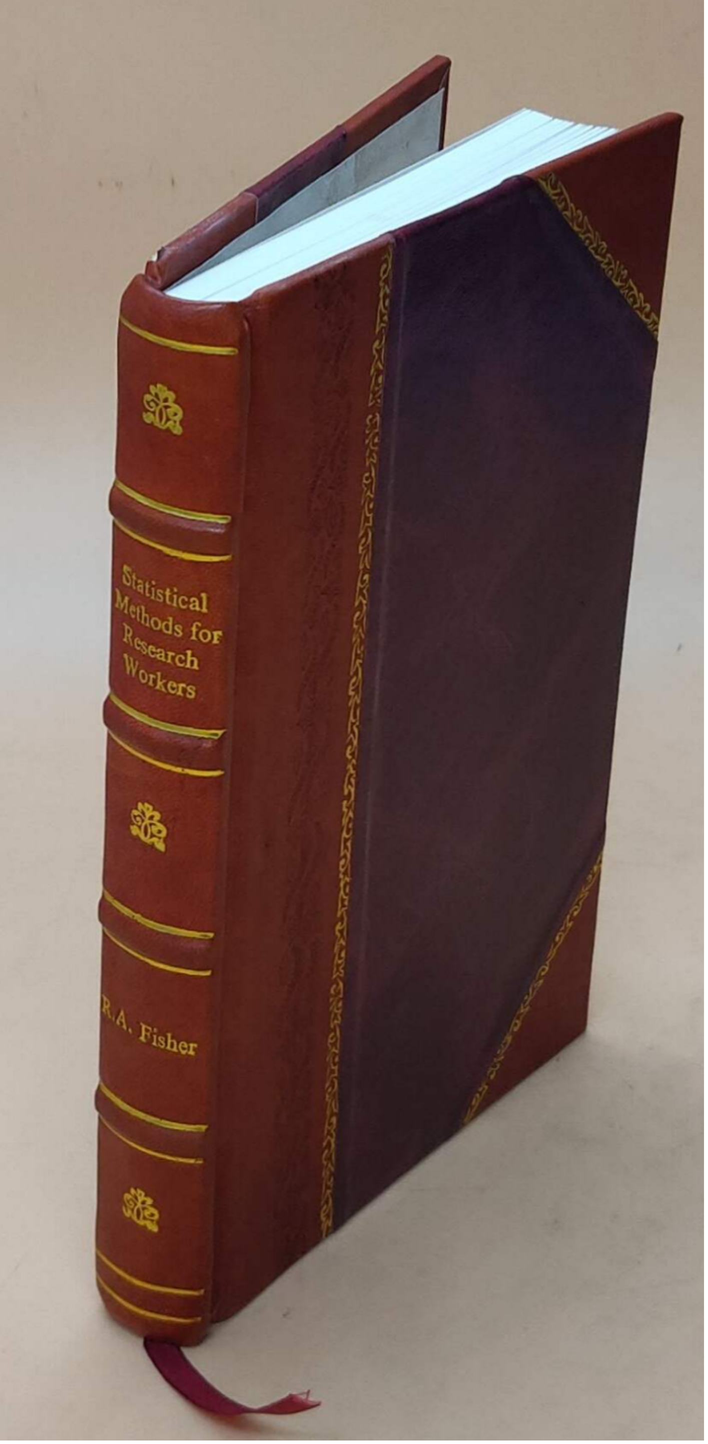


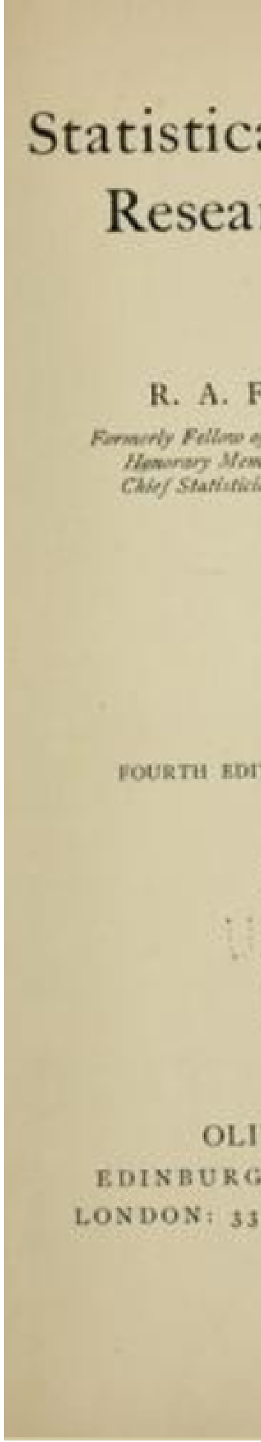
Fatal Car Accident Victims

Reported victims of fatal car accidents in 2020

Source: Forbes Advisor • Embed • Download image

In 2020, a total of 35,766 fatal motor vehicle accidents occurred on U.S. roadways.[2]
38,824 deaths.[5]



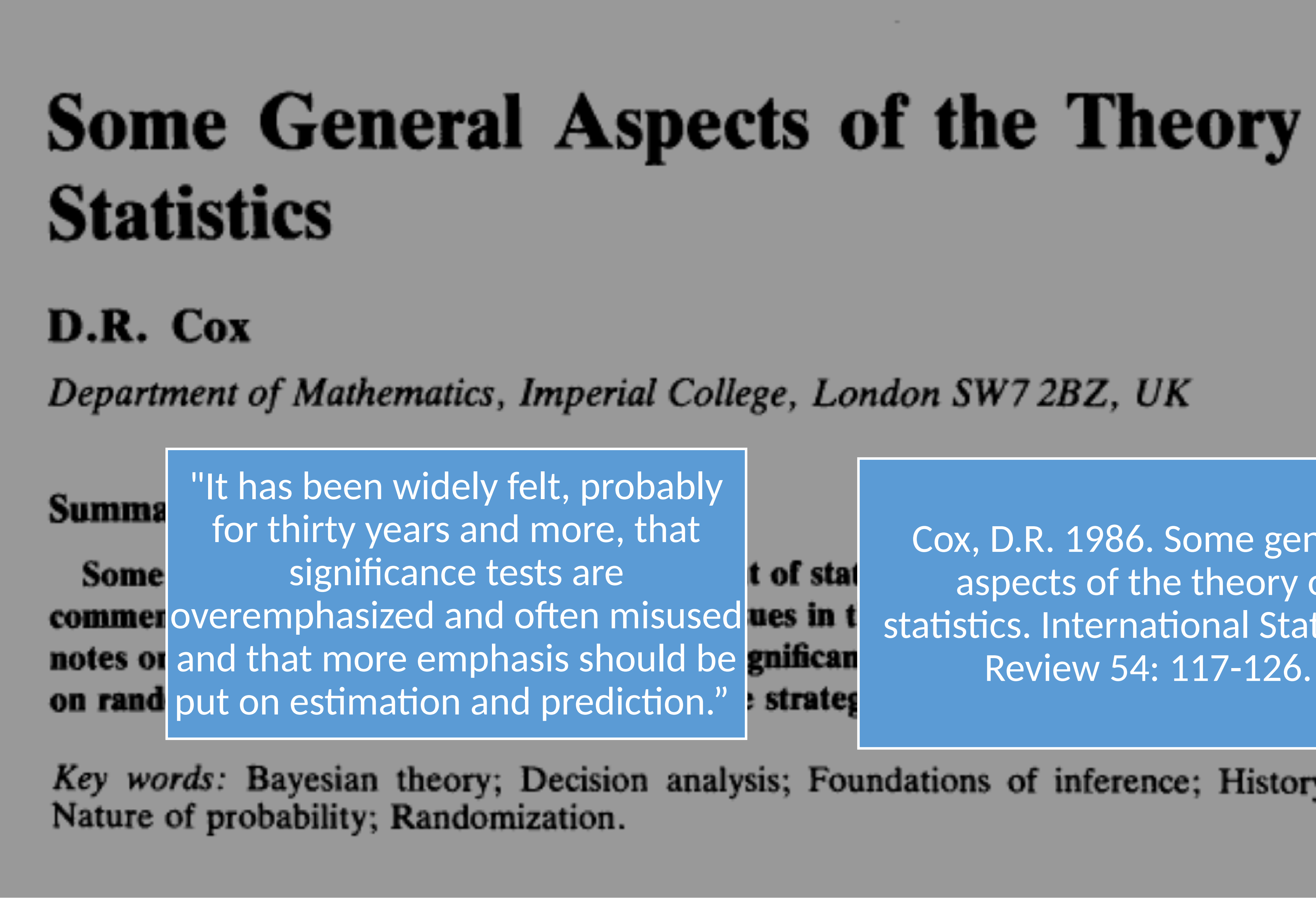


Statistica
Resea

R. A. F
Formerly Fellow o
Honorary Mem
Chief Statistici

FOURTH EDI

OLI
EDINBURG
LONDON: 33



Some General Aspects of the Theory Statistics

D.R. Cox

Department of Mathematics, Imperial College, London SW7 2BZ, UK

Summa

Some t of stat
commer ues in t
notes o gnifican
on rand strateg

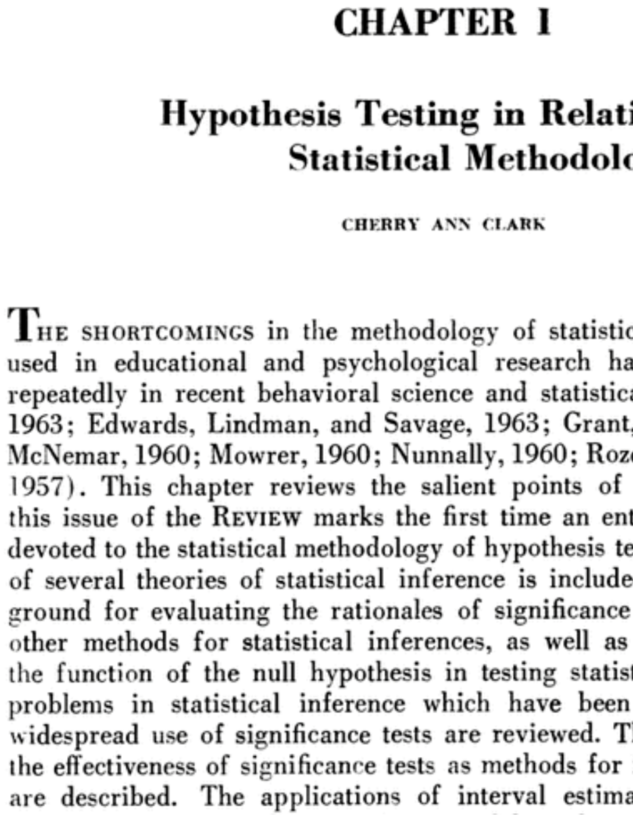
Key words: Bayesian theory; Decision analysis; Foundations of inference; Histor
Nature of probability; Randomization.

"It has been widely felt, probably for thirty years and more, that significance tests are overemphasized and often misused and that more emphasis should be put on estimation and prediction."

Cox, D.R. 1986. Some gen aspects of the theory o statistics. International Stat Review 54: 117-126.

The null hypothesis of no difference has been judged to be no longer a sound or fruitful basis for statistical investigation. […] Significance tests do not provide the information that scientists need, and, furthermore, they are not the most effective method for analyzing and summarizing data.“

- *Cherry A Clark, "Hypothesis Testing in Relation to Statistical Methodology", Review of Educational Research Vol. 33, 1963*



CHAPTER I

Hypothesis Testing in Relati
Statistical Methodol

CHERRY ANN CLARK

THE SHORTCOMINGS in the methodology of statistic
used in educational and psychological research ha
repeatedly in recent behavioral science and statistica
1963; Edwards, Lindman, and Savage, 1963; Grant,
McNemar, 1960; Mowrer, 1960; Nunnally, 1960; Roz
1957). This chapter reviews the salient points of
this issue of the REVIEW marks the first time an ent
devoted to the statistical methodology of hypothesis te
of several theories of statistical inference is include
ground for evaluating the rationales of significance
other methods for statistical inferences, as well as
the function of the null hypothesis in testing statist
problems in statistical inference which have been
widespread use of significance tests are reviewed. T
the effectiveness of significance tests as methods for
are described. The applications of interval estima

What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!

- Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

The Earth Is Round ($p < .0$

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of p *as the probability that* H_0 *is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects* H_0 *one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication.*

I make no pretense of the originality of my remarks in this article. One of the few things we, as psychologists, have learned from over a century of scientific study is that at age three score and 10, originality is not to be expected. David Bakan said back in 1966 that his claim that "a great deal of mischief has been associated" with the test of significance "is hardly original," that it is "what 'everybody knows,'" and that "to say it 'out loud' is . . . to assume the role of the child who pointed out that the emperor was really outfitted in his underwear" (p. 423). If it was hardly original in 1966, it can hardly be original now. Yet this naked emperor has been shamelessly running around for a long time.

Like many men my age, I mostly grouse. My harangue today is on testing for statistical significance, about which Bill Rozeboom (1960) wrote 33 years ago, "The statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

And today, they continue to continue. And we, as teachers, consultants, authors, and otherwise perpetrators of quantitative methods, are responsible for the ritualization of null hypothesis significance testing (NHST; I resisted the temptation to call it statistical hypothesis inference testing) to the point of meaninglessness and beyond. I argue herein that NHST has not only failed to support the advance of psychology as a science but also has seriously impeded it.

Consider the following: A colleague approaches me with a statistical problem. He believes that a generally rare disease does not exist at all in a given population, hence H_0: $P = 0$. He draws a more or less random sample of 30 cases from this population and finds that one of the cases has the disease, hence $P_s = 1/30 = .033$. He is not

sure how to test H_0, chi-sc
rection or the Fisher exact
has enough power. Would
believe that if he tried to
significance test, one or mo
It could happen.

Almost a quarter of a
ciologists, D. E. Morrison a
a book entitled *The Signific*
the contributors were Bill R
(1967), David Bakan (196
Without exception, they d
Meehl described NHST as
tual rake who leaves in hi
ravished maidens but no
265). They were, however,
Joseph Berkson attacked N
sank its deep roots in ps
book-length critique appe
then, I was appalled by its
time well trained in the cur
not yet heard of Neyman-P
to them in the statistics t
Edwards, Guilford, Walke
some dizzying success as a
NHST to my fellow clinic
tration.

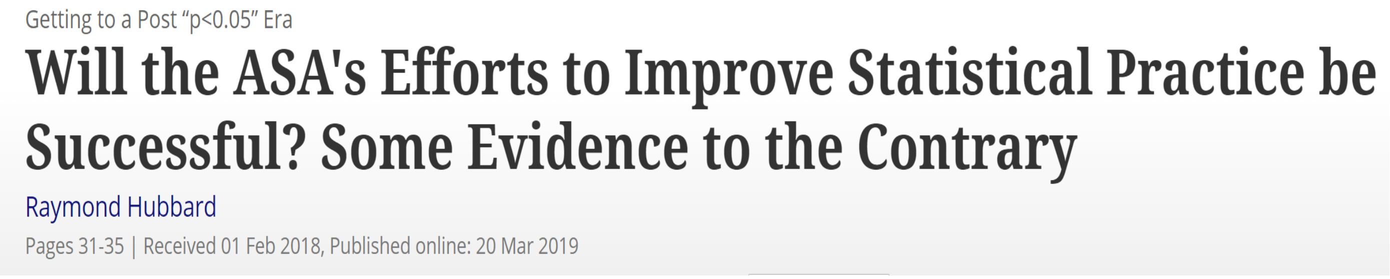
What's wrong with NH
things, it does not tell us w
so much want to know wh
of desperation, we neverthe
we want to know is "Given
ability that H_0 is true?" Bu
tells us is "Given that H_0 i
of these (or more extreme)
as has been pointed out ma
contributors to the Morriso

J. Bruce Overmier served as actio
This article was originally a
Memorial Lifetime Achievement
imental Psychology, San Pedro, C
I have made good use of th
draft of this article by Patricia C
Abelson, David Bakan, Michael
Falk, Gerd Gigerenzer, Charles Gr
ald F. Klein, Robert S. Lee, Paul
Rosenthal, William W. Rozeboon
Bruce Thompson. I also acknowle
David Lykken, Matt McGue, and
Correspondence concerning t
Cohen, Department of Psycholog
Place, 5th Floor, New York, NY

December 1994 • American Psychologist

Copyright 1994 by the American Psychological Association, Inc. 0003-066X/94/$2.00
Vol. 49, No. 12, 997–1003

At some point we should realize that more driver education is not going to do the trick!


Getting to a Post “p<0.05” Era

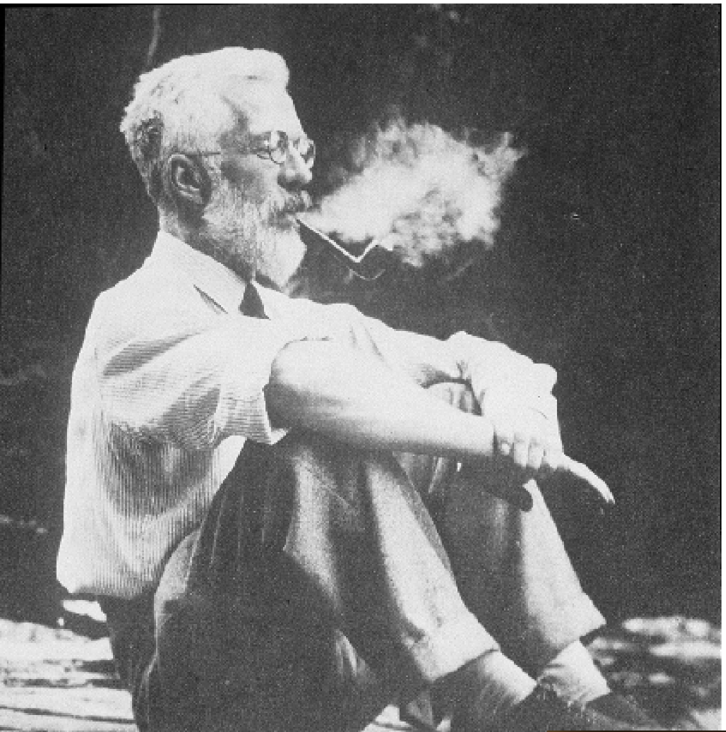
Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary

Raymond Hubbard

Pages 31-35 | Received 01 Feb 2018, Published online: 20 Mar 2019

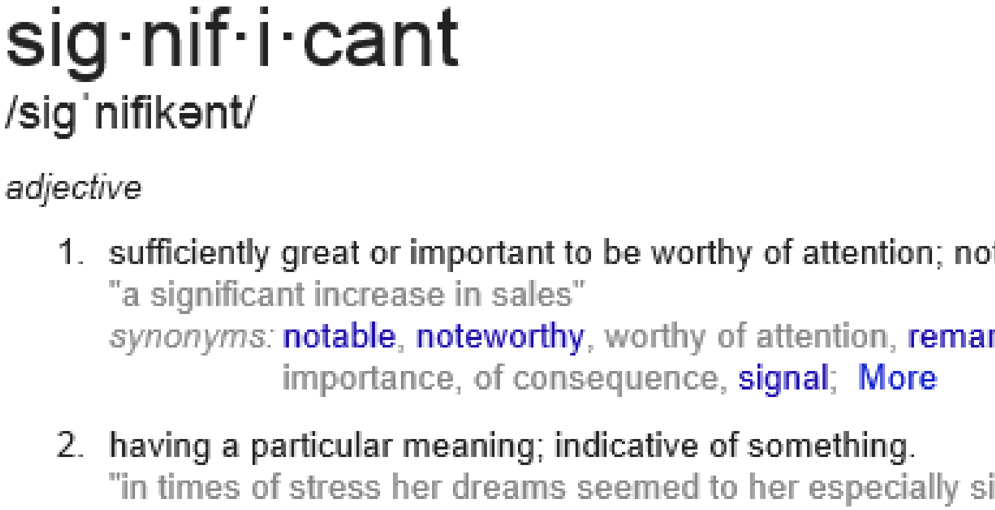
More “driver ed
has done noth
stem use and

Hubbard shows that the numb
articles critical of significance t
warning of its dangers has gro
six decades, but at the same ti
percentage of papers in many
use it has also considerably inc



R. A. Fisher called such r
“significant”

To Fisher, this meant tha
was worth further scruti

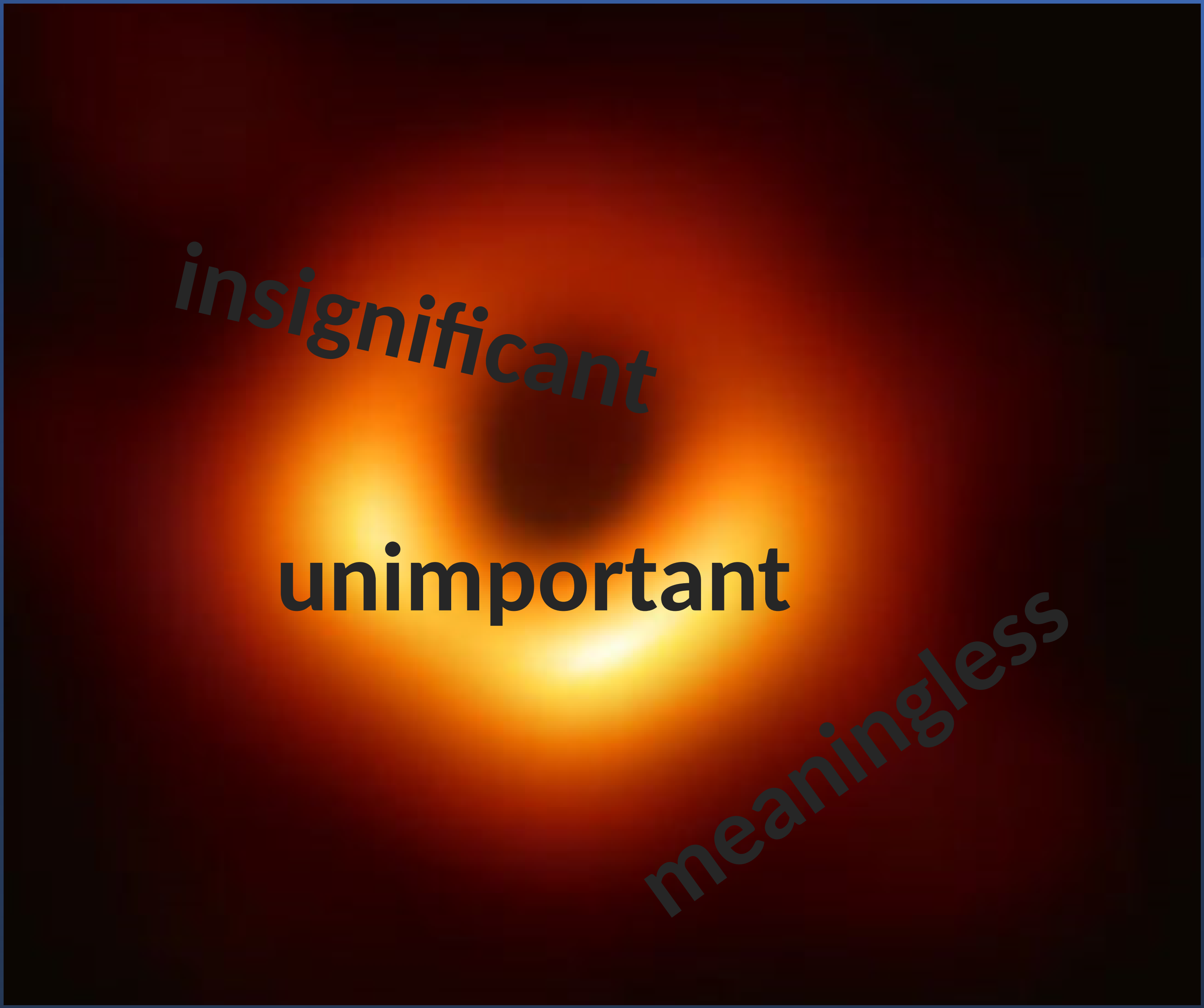


sig·nif·i·cant

/sigˈnifikənt/

adjective

1. sufficiently great or important to be worthy of attention; no
   "a significant increase in sales"
   synonyms: notable, noteworthy, worthy of attention, remar
   importance, of consequence, signal; More
2. having a particular meaning; indicative of something.
   "in times of stress her dreams seemed to her especially si



insignificant

unimportant

meaningless

significant increase

significant event

significant other

mole

The amount or sample of a chemical substance that contains as many constitutive particles, e.g., atoms, molecules, ions, electrons, or photons, as there are atoms in 12 grams of carbon-12

“You keep using that word. I don’t think that means what you think it means.” – Inigo Mont

“Just a Theory”: 7 Misused Scientific Words, Scientific American, April 2, 2013 https://www.scientificamerican.com/article/just-a-theory-7-misused-science-words/

Word #1

Hypothesis

A proposed explanation **that can be tested**

Word #2

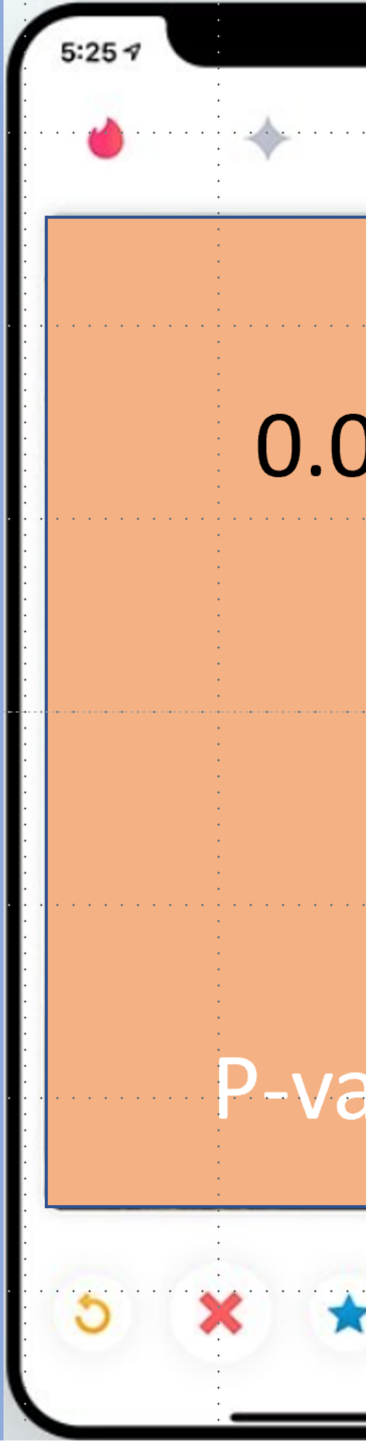
Theory

An explanation of some aspect of the natural world that has been **substantiated through repeated experiments or testing**

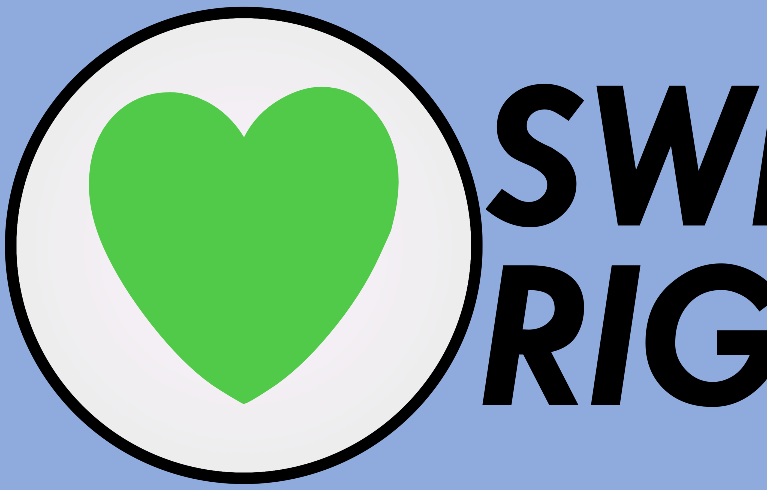
Word #6

Significant

tinder™

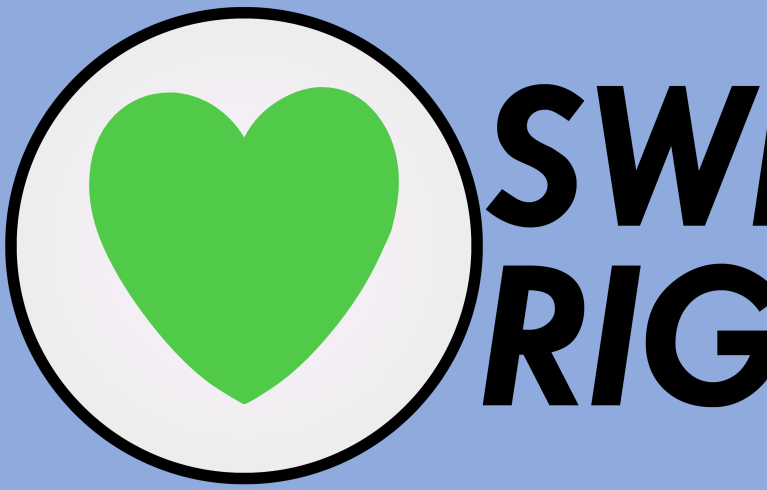


My experimental results are interesting. I should spend more time with them, maybe repeat the experiment. I may be on to something, but it will take time to be sure.

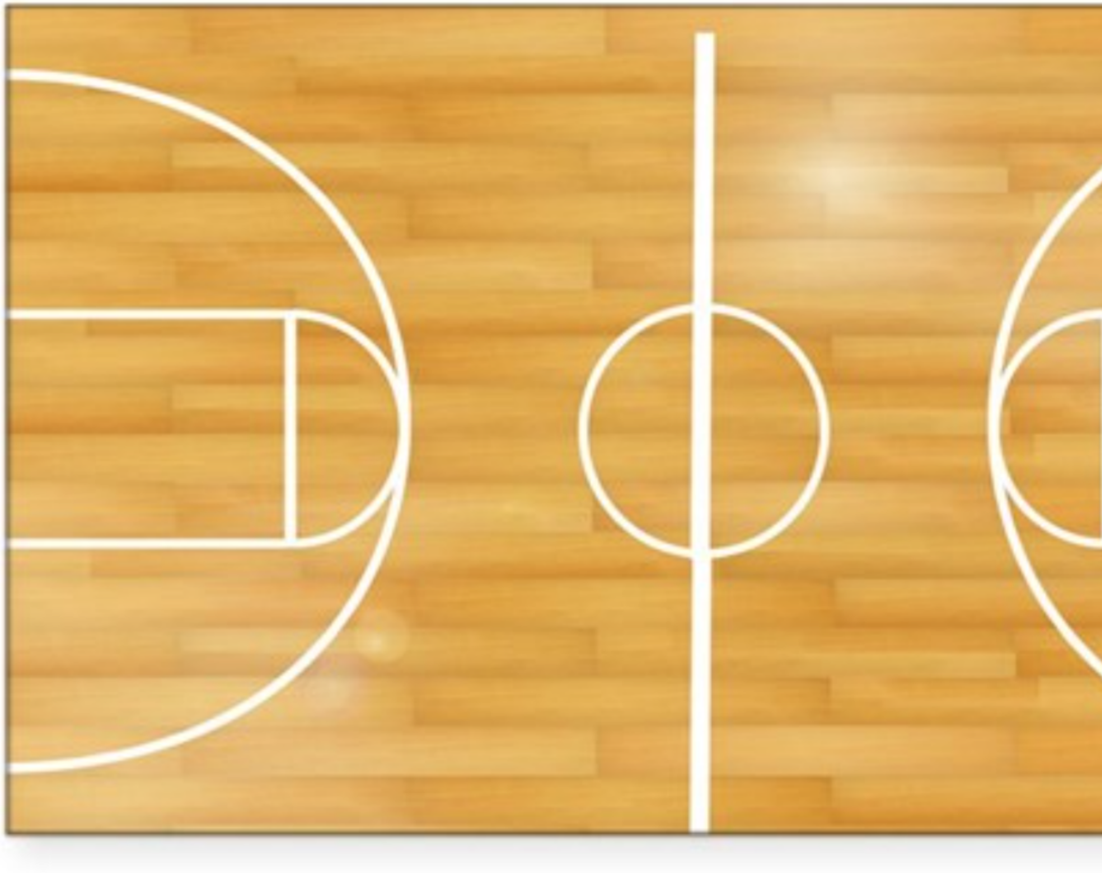


You tiny, beautiful p-value. You are the result that I want to spend the rest of my life with. Let’s publish and get grants together.

I love you!



A word about thresh olds



Boundary lines

Boundary	Arbitrary	Rational
Necessary		
Unnecessary		

Boundary lines

Boundary	Arbitrary	Rational
Necessary	Soccer	
Unnecessary		

Boundary lines

Boundary	Arbitrary	Rational
Necessary	Soccer	Property
Unnecessary		

Boundary lines

Boundary	Arbitrary	Rational
Necessary	Soccer	Property
Unnecessary		Traffic lanes certain cou

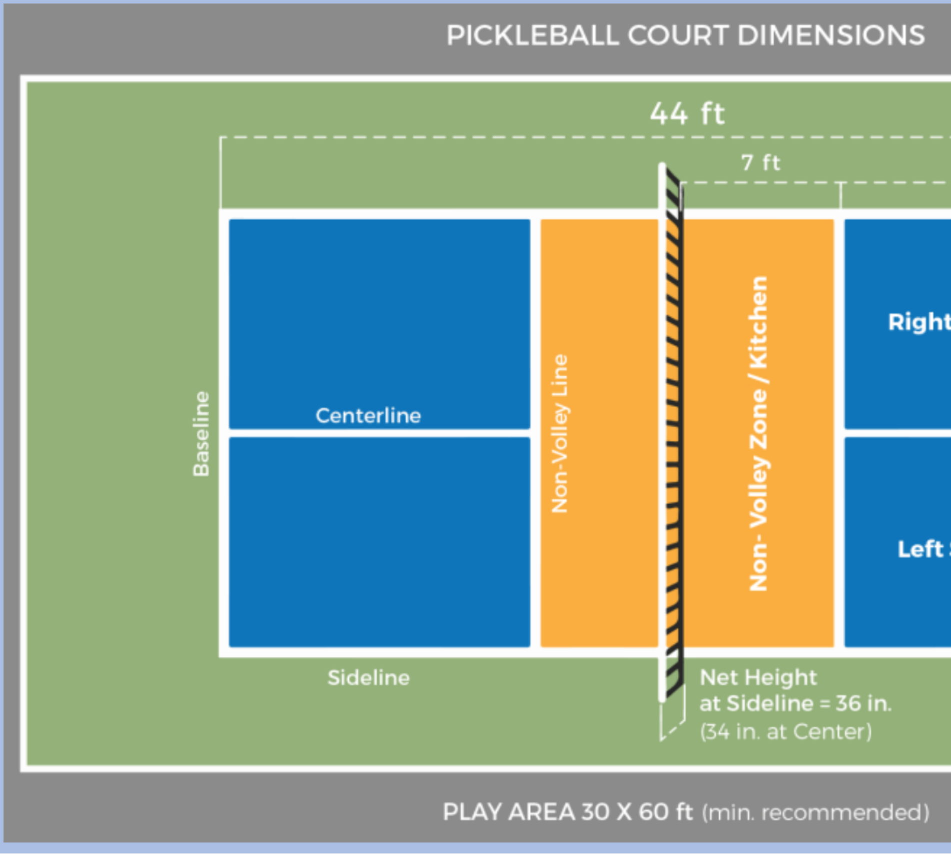
Boundary lines

Boundary	Arbitrary	Rational
Necessary	Soccer	Property
Unnecessary	$p < 0.05$	Traffic lanes certain cou

Boundary lines in sports

The ball is still in-bounds if it touches the line in

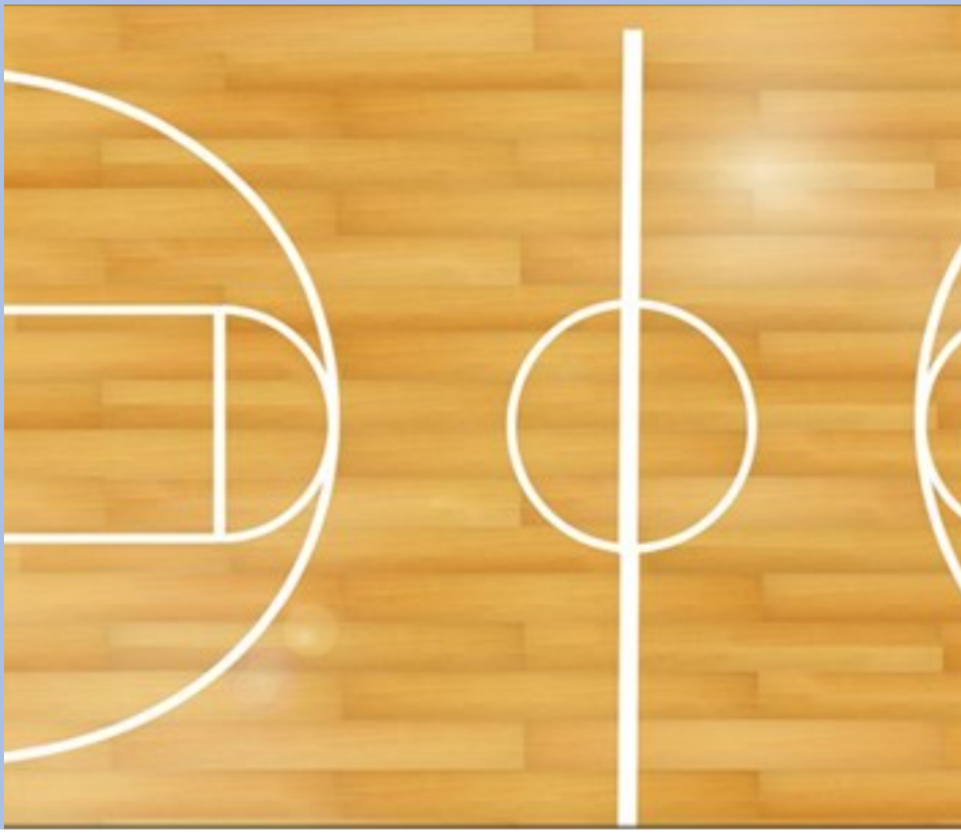
- Baseball
- Tennis
- Soccer
- Volleyball
- Pickleball



Boundary lines in sports

The ball is out-of-bounds if it touches the line in

- Football
- Basketball



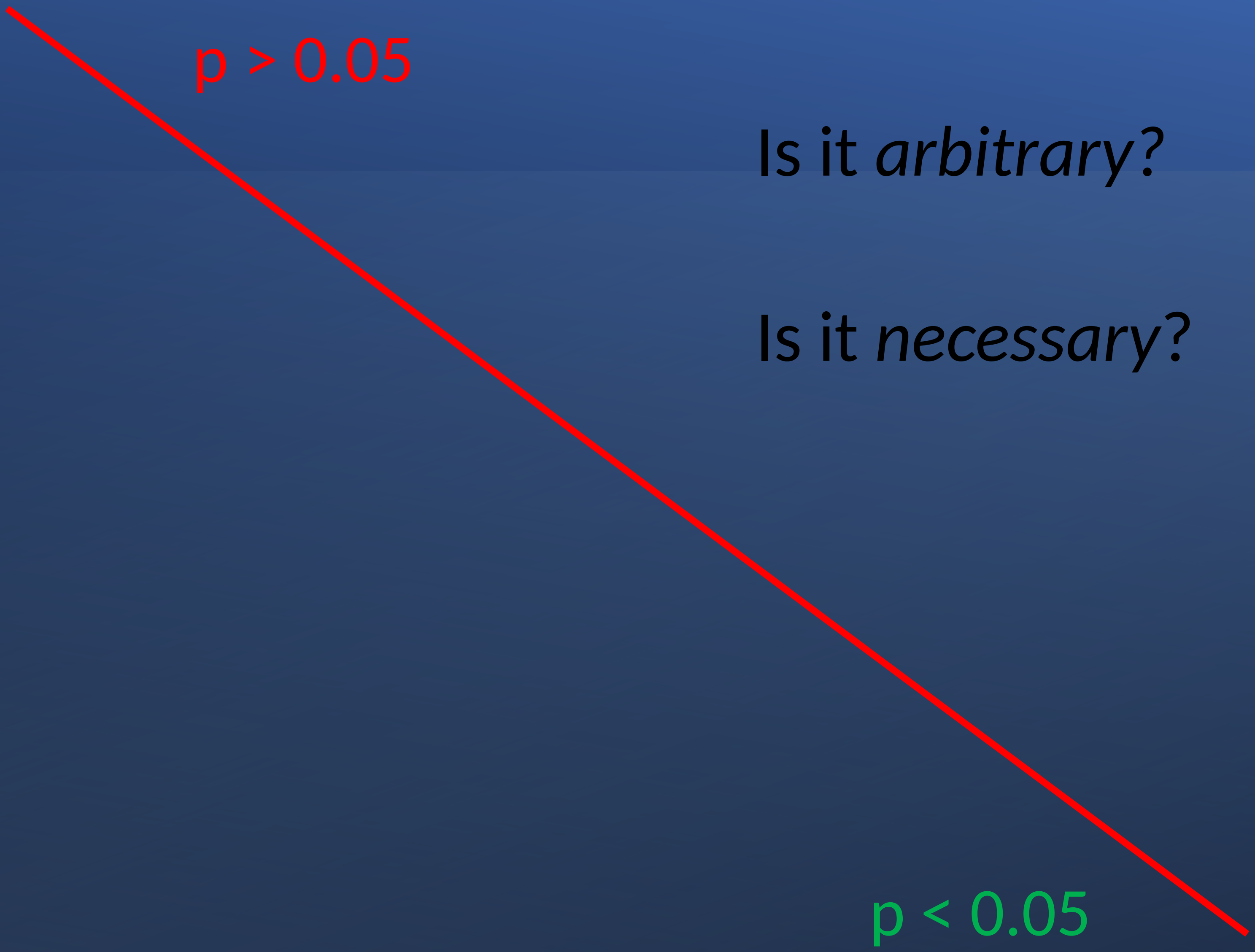
These boundaries are integral to the play of game

- Landing outside the boundary produ very different outcome than landing
- They are *arbitrary but necessary* bou
  - Arbitrary: established over time b history of the sport or the size of playing area (soccer: 90-120m x 4
  - Necessary: The game needs the boundaries to regulate play

How about this boundary

Is it *arbitrary*?

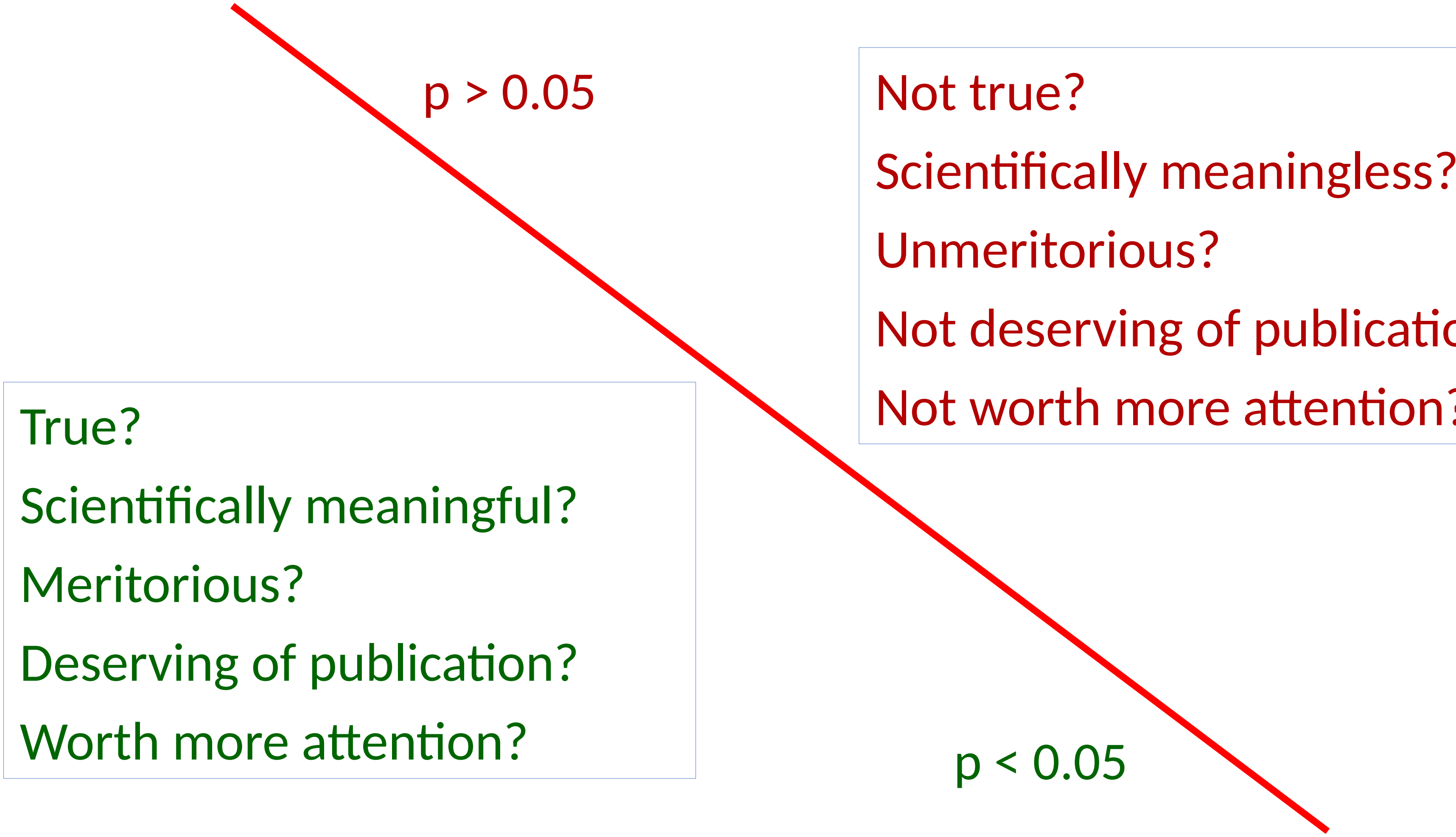
Is it *necessary*?



$p > 0.05$

$p < 0.05$

Landing outside THIS boundary also produ
very different outcome than landing inside



p > 0.05

Not true?

Scientifically meaningless?

Unmeritorious?

Not deserving of publicatio

Not worth more attention?

True?

Scientifically meaningful?

Meritorious?

Deserving of publication?

Worth more attention?

p < 0.05

THIS boundary is *arbitrary*, but it is *unnecessary*

p > 0.05

Arbitrary:

The bound
represent
perspectiv
indicates “
a convenie
history.

Unnecessary:

“Significant” does not mean that an observed effect is not due to chance. It also does not mean that the effect is real, genuine, important, true, or any of the other common misinterpretations.

p < 0.05

THIS boundary is *arbitrary*, but it is *unnecessary*

p > 0.05

Declaration of
significance is u
ending point ar
starting point -
unreliable resu
unwarranted c

A declaration of statistical significance does not convey anything useful beyond what is conveyed by the p-value itself. It adds no new evidence.

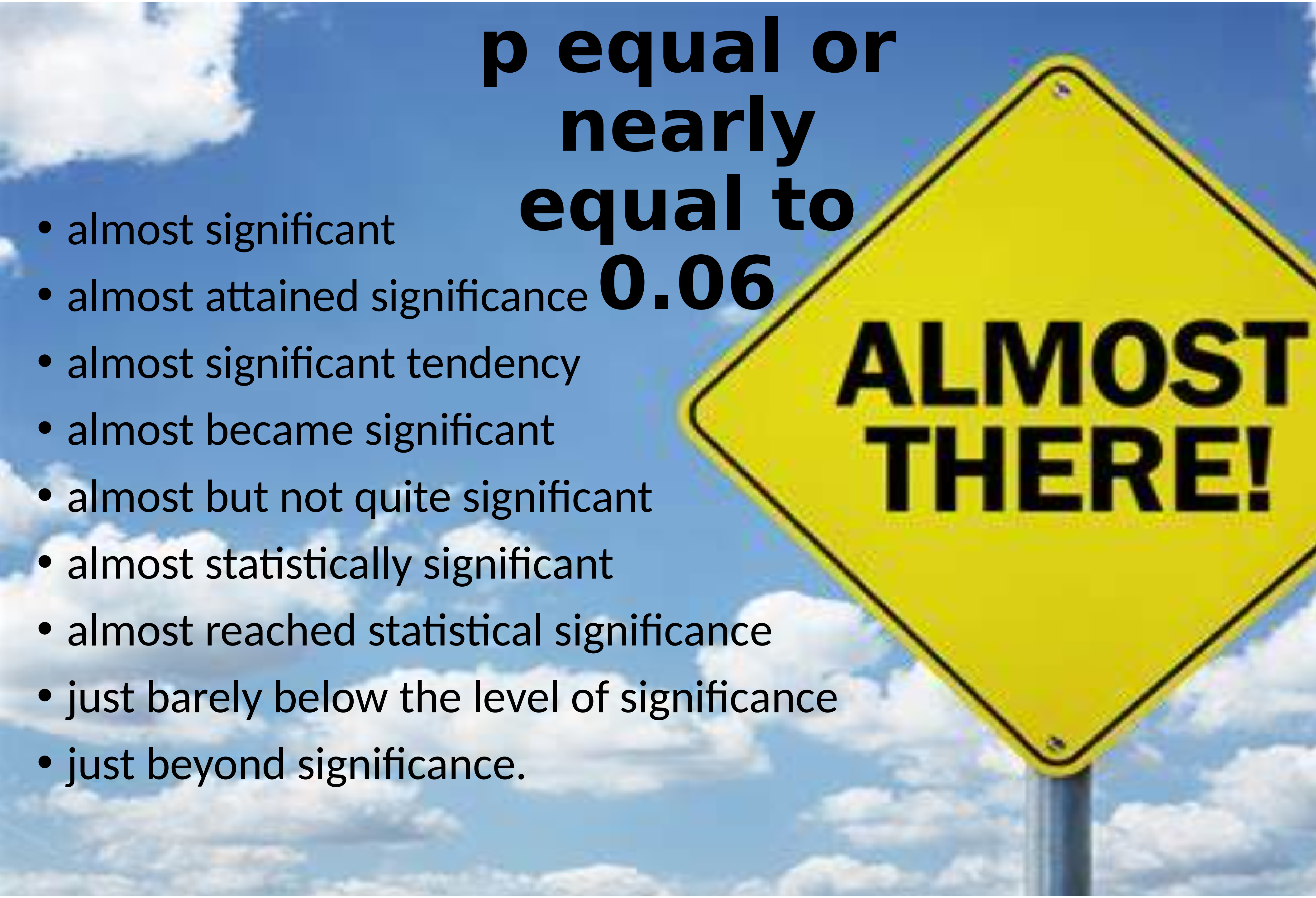
p < 0.0

Bright line thinking

- The problem with using bright lines is that they inevitably lead to our treating results on opposite sides of the line very differently, even if their practical implications are identical.
- Moreover, having such a rule establishes how to achieve a desired outcome by manipulation, and unfortunately, once achieved, that result usually gains more weight than is deserved.

"… we only wish to emphasize that dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05."

Rosnow, R.L. and Rosenthal, R. 1989. Statistical procedures and the justification of knowledge and psychological science. American Psychologist 44: 1276-1284



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.

p equal or nearly equal to 0.08

- a certain trend toward significanc
- a definite trend
- a slight tendency toward significa
- a strong trend toward significance
- a trend close to significance
- an expected trend
- approached our criteria of signific
- approaching borderline significanc
- approaching, although not reachi significance.

p close to but not less than 0.05

- hovered at nearly a significant lev
- hovers on the brink of significance
- just about significant (p=0.051)
- just above the margin of significan
- just at the conventional level of si (p=0.05001)
- just barely statistically significant (
- just borderline significant (p=0.05
- just escaped significance (p=0.057
- just failed significance (p=0.057).

Thanks to Matthew Hankin for these quotes

https://mchankins.wordpress.com/2013/04/21/still-not-significant-

time to move o

It’s time to say farewell to “statistically signif

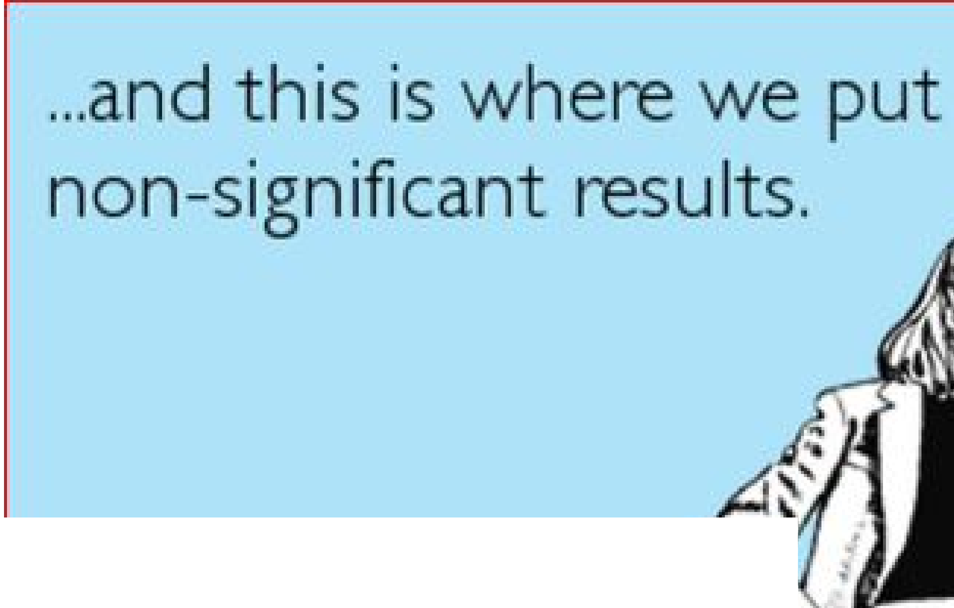
“Moving to a World Beyond p<0.05”

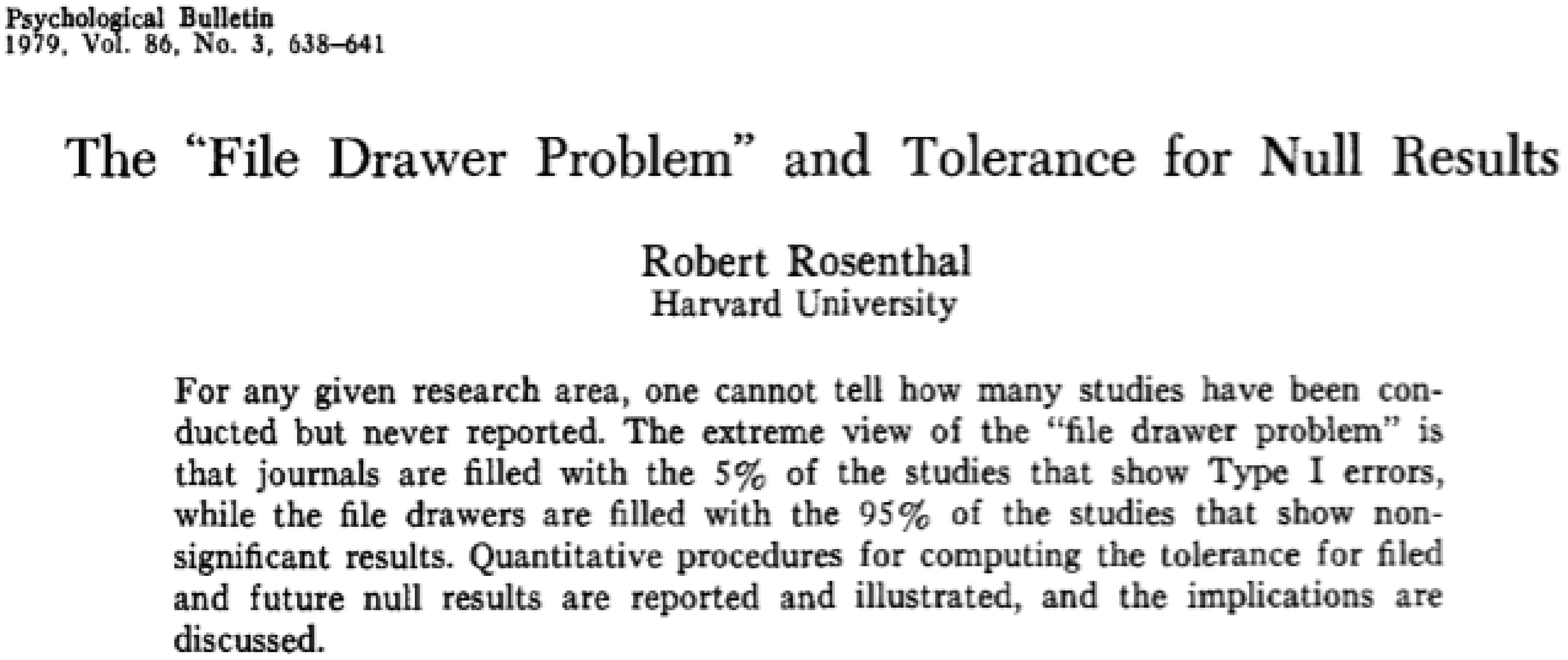
https://amstat.tandfonline.com/doi/full/10.1080/00031305.2019.1583913#.XYjKQ25FxPY

“Scientists rise up against statistical significance”

https://www.nature.com/articles/d41586-019-00857-9

- Significance has lost its
- Bright lines lead to biza
- Decades of complainin nothing
- “A label of statistical sig adds nothing to what is conveyed by the value o this dichotomization of makes matters worse.” editorial)
- Multiple analyses
- File drawer effect





Psychological Bulletin
1979, Vol. 86, No. 3, 638–641

The "File Drawer Problem" and Tolerance for Null Results

Robert Rosenthal
Harvard University

For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results. Quantitative procedures for computing the tolerance for filed and future null results are reported and illustrated, and the implications are discussed.

http://datac
content/upl
Rosenthal-1
problem-and
results.pdf

Change is needed...

...but change is never easy

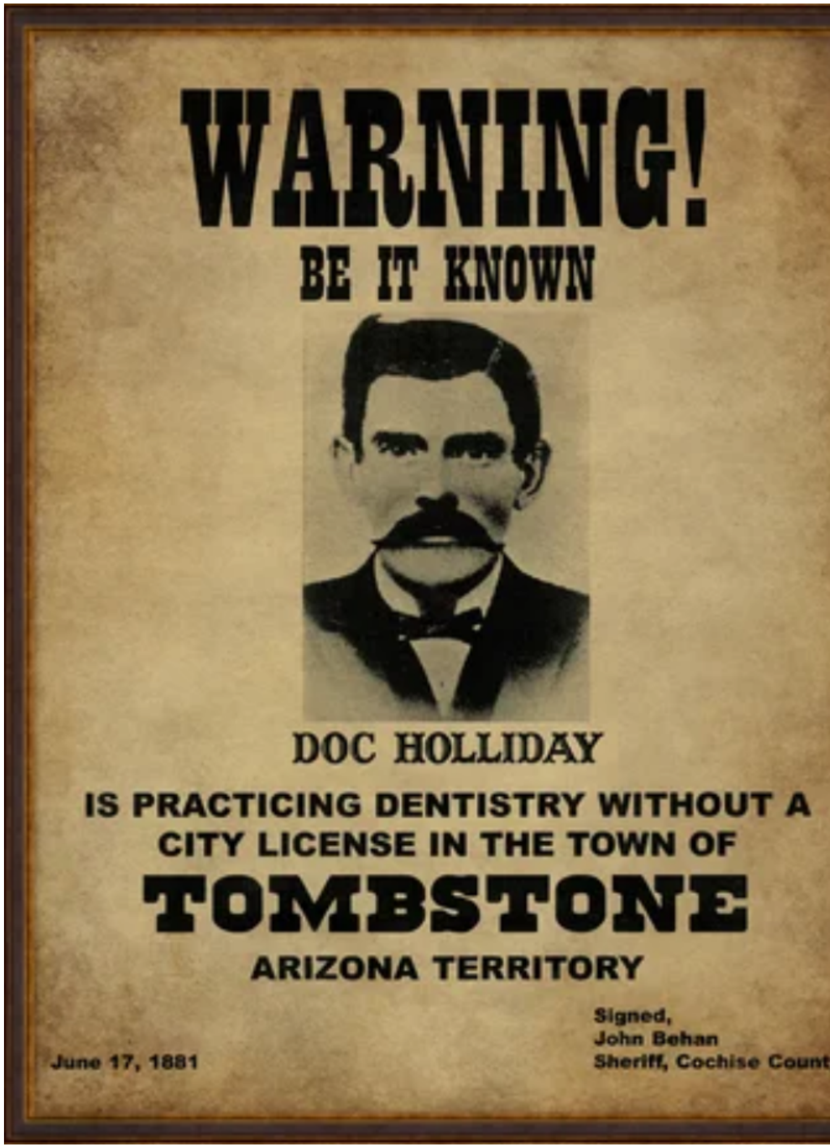
"The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It's the same reason we can use money. When everyone believes in something's value, we can use it for real things; money for food, and p-values for knowledge claims, publication, funding, and promotion. It doesn't matter if the p-value doesn't mean what people think it means; it becomes valuable because of what it buys." (Goodman – 2019 (TAS))

Before listing some changes, though, let’s be sure to note that there are…

Opposing views

1. Potentially creates anarchy
2. Negatively impacts image of s
3. Why pick on p-values?
4. Decisions have to be made

1. Ending Significance Creates “The Wild West”



WARNING!

BE IT KNOWN

DOC HOLLIDAY

IS PRACTICING DENTISTRY WITHOUT A CITY LICENSE IN THE TOWN OF

TOMBSTONE

ARIZONA TERRITORY

June 17, 1881

Signed,
John Behan
Sheriff, Cochise Count

VIEWPOINT

The Importance of Predefined Ru and Prespecified Statistical Analys

Do Not Abandon Significance

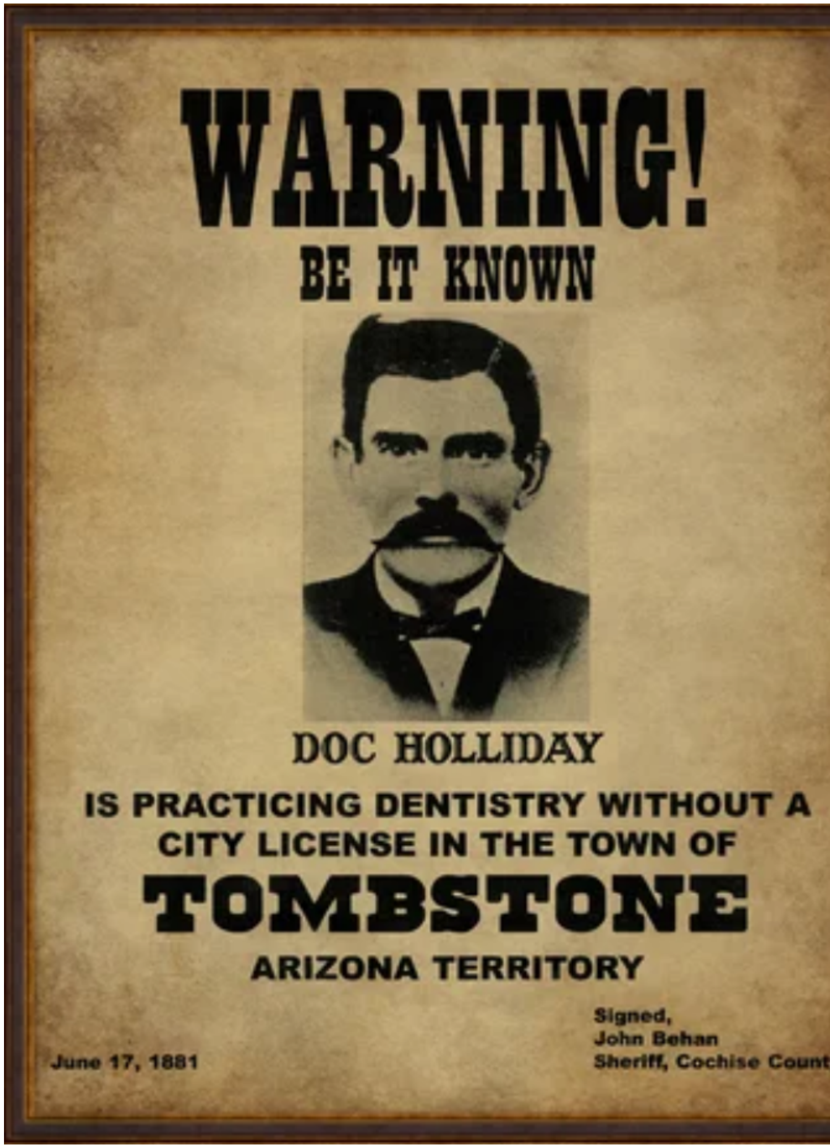
John P. A. Ioannidis, MD, DSc
Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California; and Meta-Research Innovation Center–Berlin (METRIC-B), Berlin, Germany.

The statistical numeracy of the scientific workforce requi provement. Banning statistical significance while retainin ues (or confidence intervals) will not improve numeracy and m ter statistical confusion and create problematic issues with interpretation, a state of statistical anarchy. Uniformity in s cal rules and processes makes it easier to compare like with l avoid having some associations and effects be more privilege others in unwarranted ways. Without clear rules for the analys ence and policy may rely less on data and evidence and more c jective opinions and interpretations.

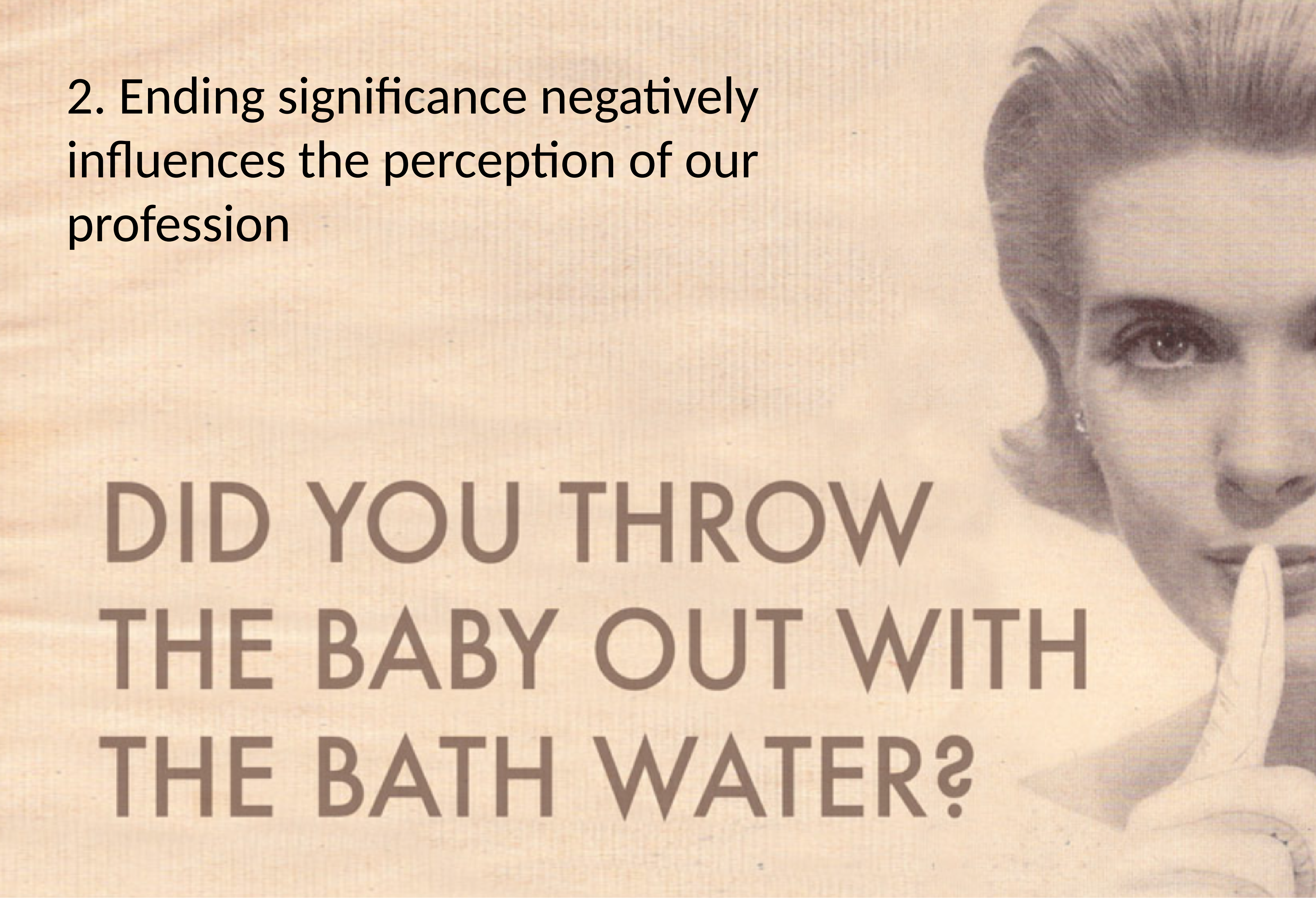
JAMA Published online Apr

This argument does not address any of the shortcomings of the use of statistical significance.

Is the way to avoid “statistical anarchy” by using a problematic method?



2. Ending significance negatively influences the perception of our profession

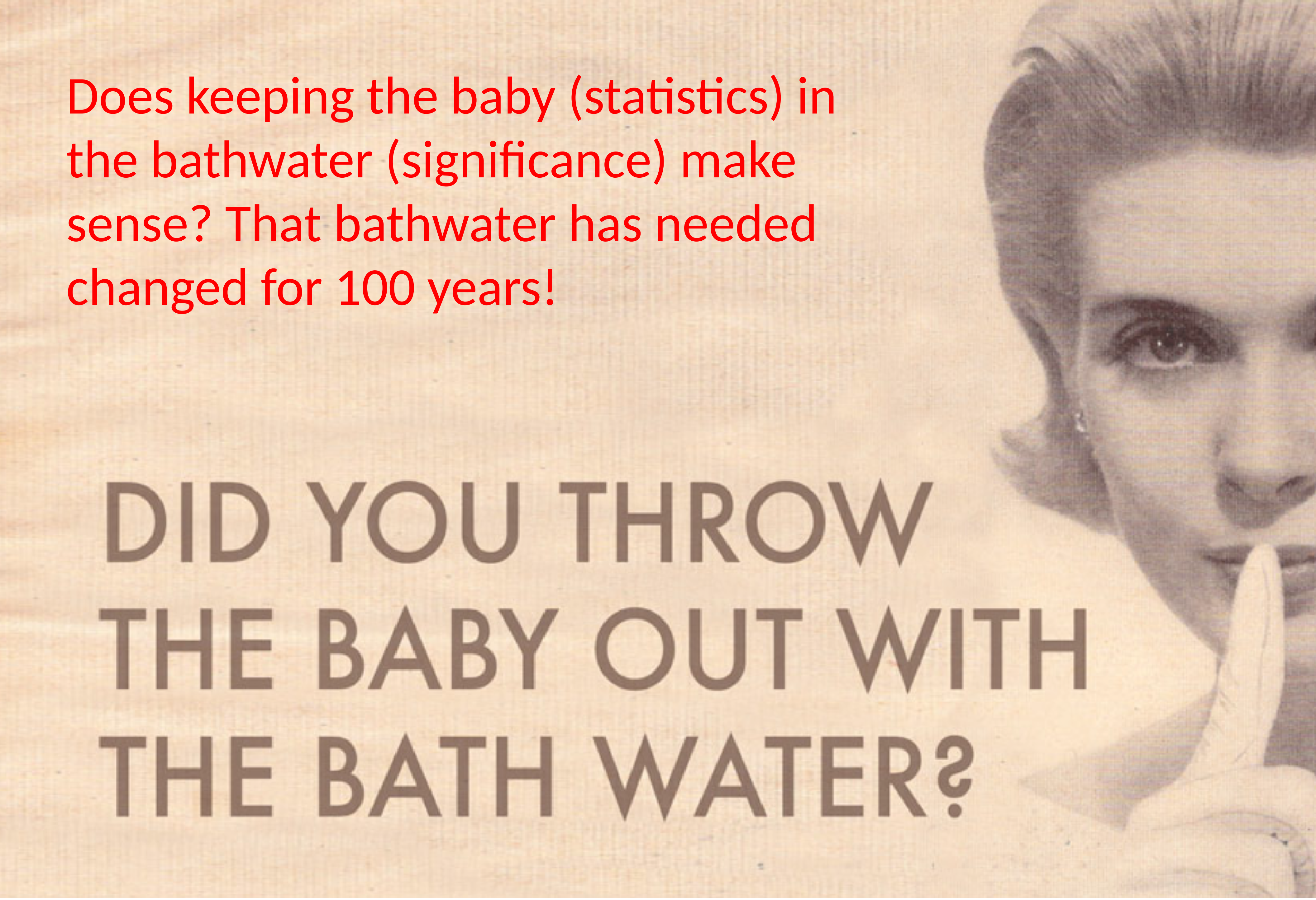


ASA President's Corner

- "…researchers may read the call t
  'abandon statistical significance' a
  'abandon statistical methods altog

- https://magazine.amstat.org/blog/2019/06/01/u
  nsequences/

Does keeping the baby (statistics) in the bathwater (significance) make sense? That bathwater has needed changed for 100 years!



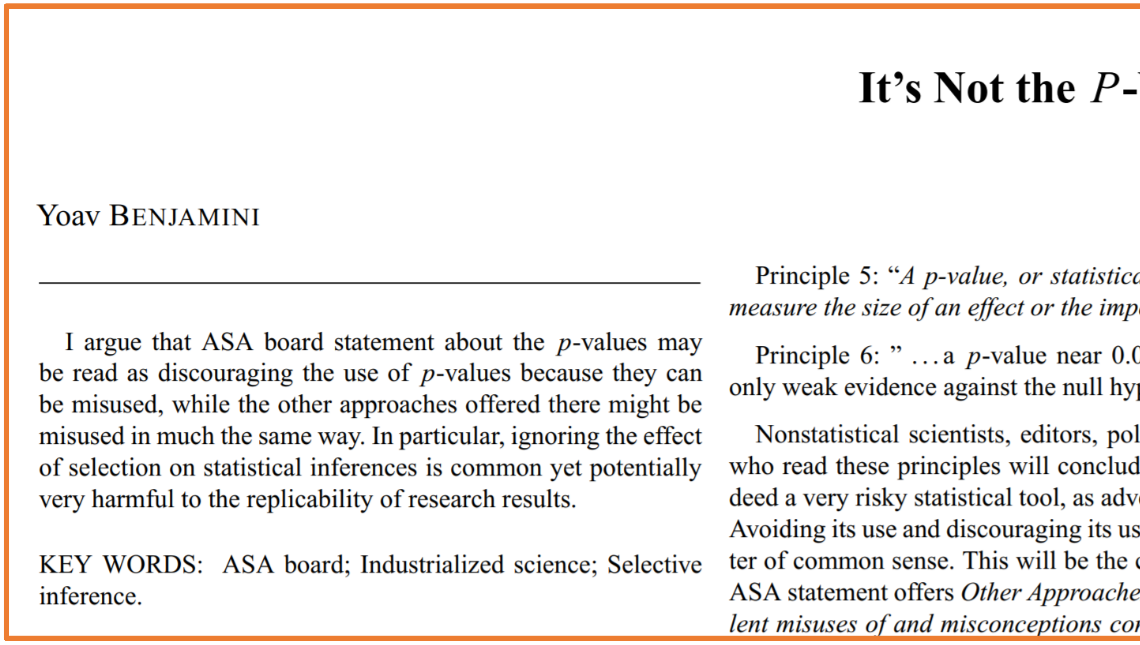
“It’s the Same Old S♫g”

3. Everything we are saying about statistical significance could be true for other statistical measures as well.

Other methods have the same problems

Benjamini, Y. Online discussion of the ASA Statement on Sta
Values, The American Statistician, 70.

“Yet all of these other approaches, as well as most statistical tools, may suffer from many of the same problems as the p-values do. What level of likelihood ratio in favor of the research hypothesis will be acceptable to the journal? Should scientific discoveries be based on whether posterior odds pass a specific threshold (P3)? Does either measure the size of an effect (P5)?”



It’s Not the P-

Yoav BENJAMINI

I argue that ASA board statement about the p-values may be read as discouraging the use of p-values because they can be misused, while the other approaches offered there might be misused in much the same way. In particular, ignoring the effect of selection on statistical inferences is common yet potentially very harmful to the replicability of research results.

KEY WORDS: ASA board; Industrialized science; Selective inference.

Principle 5: “*A p-value, or statistica*
measure the size of an effect or the imp

Principle 6: ” . . . a p-value near 0.0
only weak evidence against the null hyp

Nonstatistical scientists, editors, pol
who read these principles will conclud
deed a very risky statistical tool, as adv
Avoiding its use and discouraging its us
ter of common sense. This will be the c
ASA statement offers *Other Approache*
lent misuses of and misconceptions co

“It’s the Same Old S♫g”

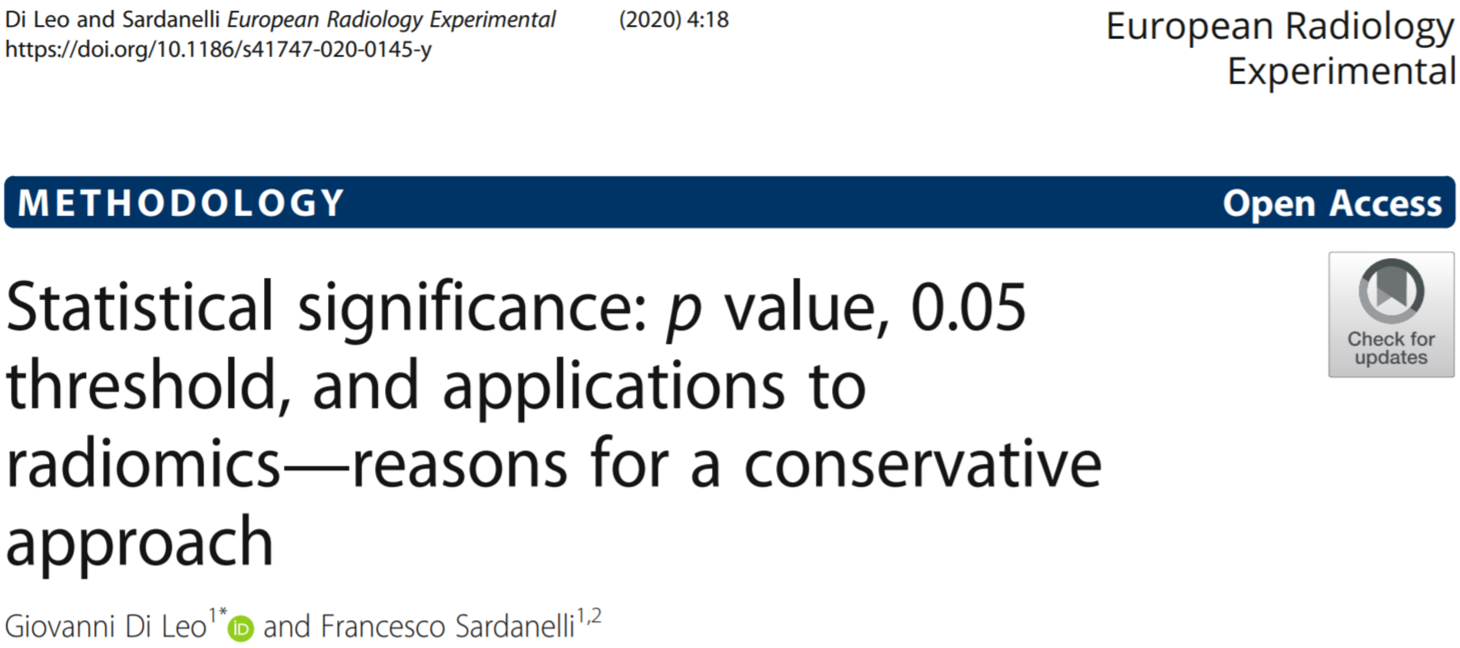
True!

But that doesn’t imply that we should keep using a method that we KNOW has been abused for decades because other methods could be similarly abused.

4. Decisions have to be made

The Clash - Should I Stay or Should I Go (Official Aud

YouTube · 120.8M views · Aug 8, 2016



Di Leo and Sardanelli *European Radiology Experimental* (2020) 4:18
https://doi.org/10.1186/s41747-020-0145-y

European Radiology
Experimental

METHODOLOGY Open Access

Check for
updates

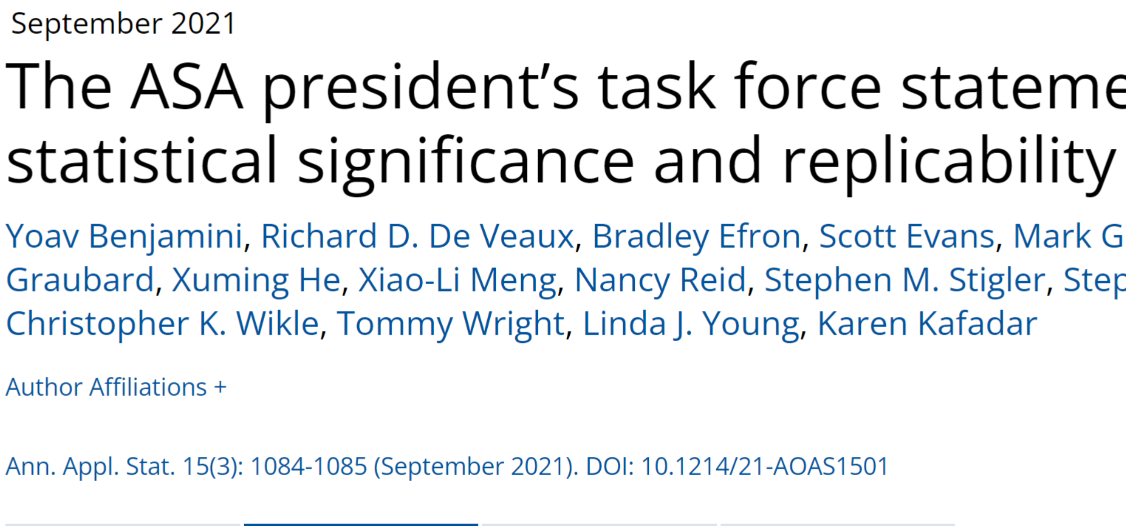
Statistical significance: *p* value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach

Giovanni Di Leo[1*] and Francesco Sardanelli[1,2]

We acknowle
importance
embracing u
avoiding hyp
and recogni
the *p* value i
poorly under
statistical si
in our opinio
crucial pract
importance.
are unavoida
dichotomous
especially in
and healthca
in preclinica
(*experiment*
clinical rese

ASA Task Force statement

“Its purpose is two-fold: to clarify that the use of P -values and significance testing, properly applied and interpreted, are important tools that should not be abandoned, and to briefly set out some principles of sound statistical inference that may be useful to the scientific community.”



September 2021

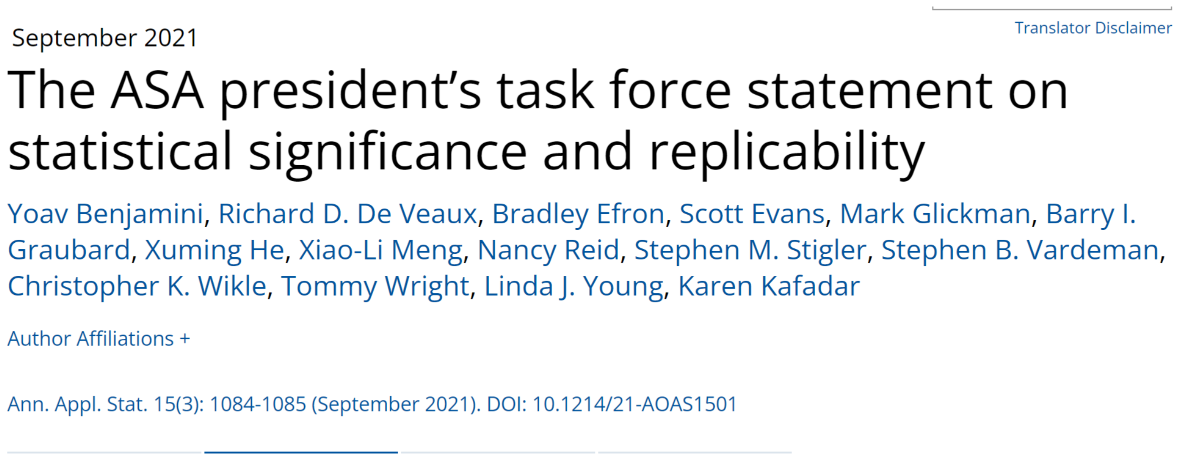
The ASA president’s task force stateme
statistical significance and replicability

Yoav Benjamini, Richard D. De Veaux, Bradley Efron, Scott Evans, Mark G
Graubard, Xuming He, Xiao-Li Meng, Nancy Reid, Stephen M. Stigler, Step
Christopher K. Wikle, Tommy Wright, Linda J. Young, Karen Kafadar

Author Affiliations +

Ann. Appl. Stat. 15(3): 1084-1085 (September 2021). DOI: 10.1214/21-AOAS1501

“**Thresholds are helpful when actions a required.** Comparing P-values to a signi level can be useful…. If thresholds are d necessary as a part of decision-making, should be explicitly defined based on st goals, considering the consequences of decisions. Conventions vary by disciplin purpose of analyses.” [highlighting in original]



Translator Disclaimer

September 2021

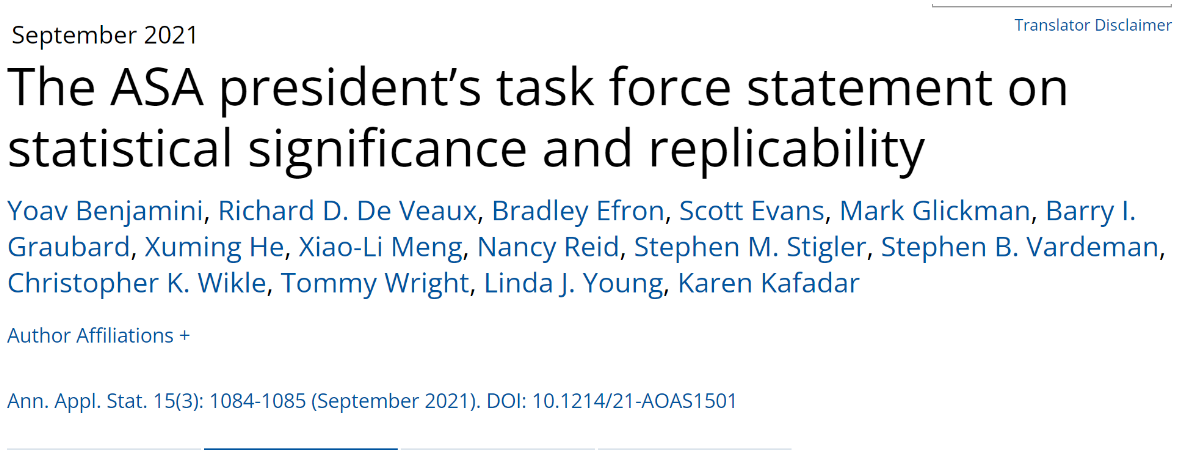
The ASA president’s task force statement on statistical significance and replicability

Yoav Benjamini, Richard D. De Veaux, Bradley Efron, Scott Evans, Mark Glickman, Barry I. Graubard, Xuming He, Xiao-Li Meng, Nancy Reid, Stephen M. Stigler, Stephen B. Vardeman, Christopher K. Wikle, Tommy Wright, Linda J. Young, Karen Kafadar

Author Affiliations +

Ann. Appl. Stat. 15(3): 1084-1085 (September 2021). DOI: 10.1214/21-AOAS1501

“**Thresholds are helpful when actions a**
required. Comparing P-values to a signi
level can be useful.... If thresholds are d
necessary as a part of decision-making,
should be explicitly defined based on st
goals, considering the consequences of
decisions. Conventions vary by disciplin
purpose of analyses.” [highlighting in
original]



Translator Disclaimer

September 2021

The ASA president’s task force statement on statistical significance and replicability

Yoav Benjamini, Richard D. De Veaux, Bradley Efron, Scott Evans, Mark Glickman, Barry I. Graubard, Xuming He, Xiao-Li Meng, Nancy Reid, Stephen M. Stigler, Stephen B. Vardeman, Christopher K. Wikle, Tommy Wright, Linda J. Young, Karen Kafadar

Author Affiliations +

Ann. Appl. Stat. 15(3): 1084-1085 (September 2021). DOI: 10.1214/21-AOAS1501

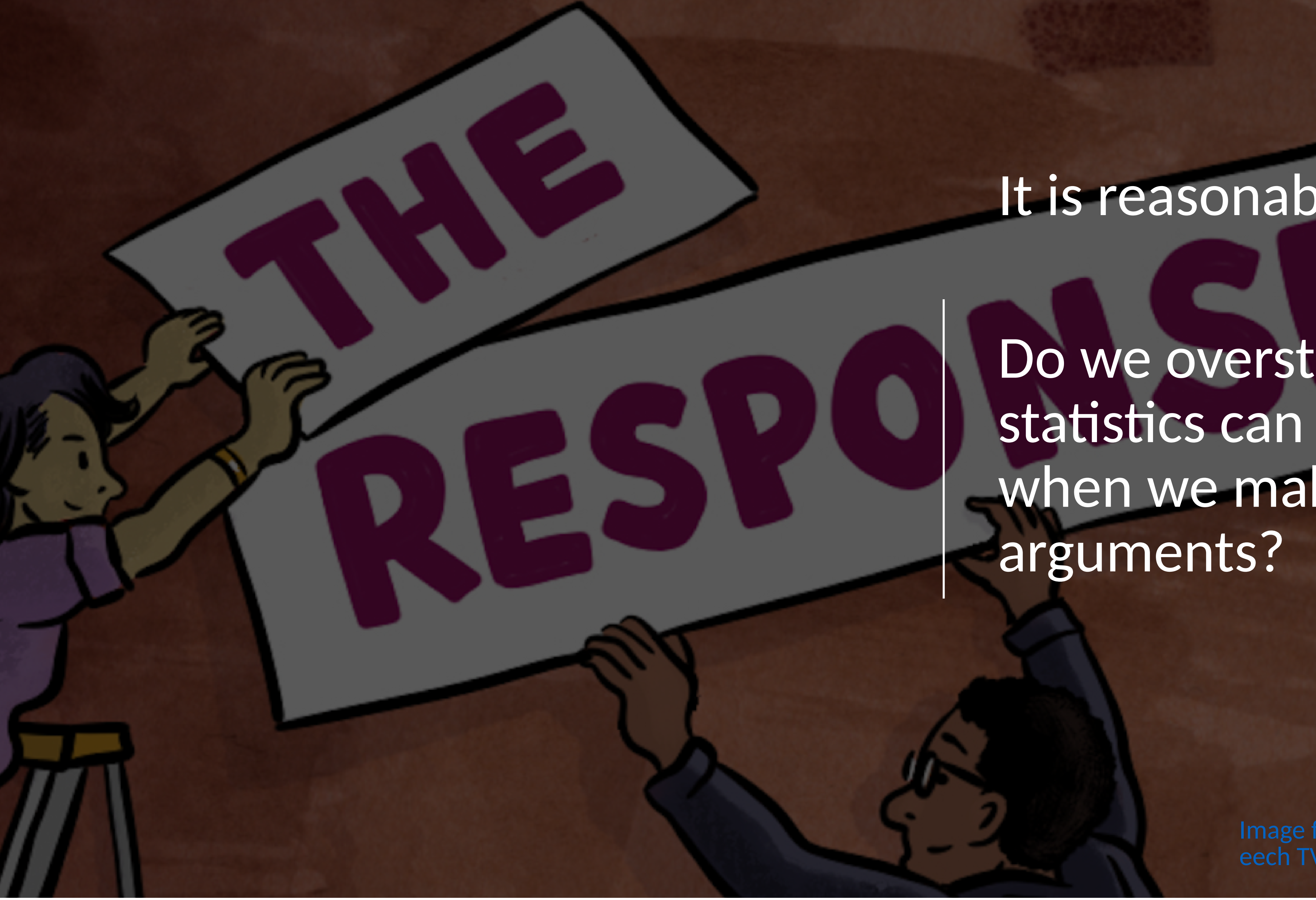
4. Decisions have to be made

The Clash - Should I Stay or Should I Go (Official Auc

YouTube · 120.8M views · Aug 8, 2016

Decisions might be dichotomous. But strength of evidence is not.

And though we know thresholds "should be explicitly defined based on study goals, considering the consequences of incorrect decisions," that's not what researchers do.



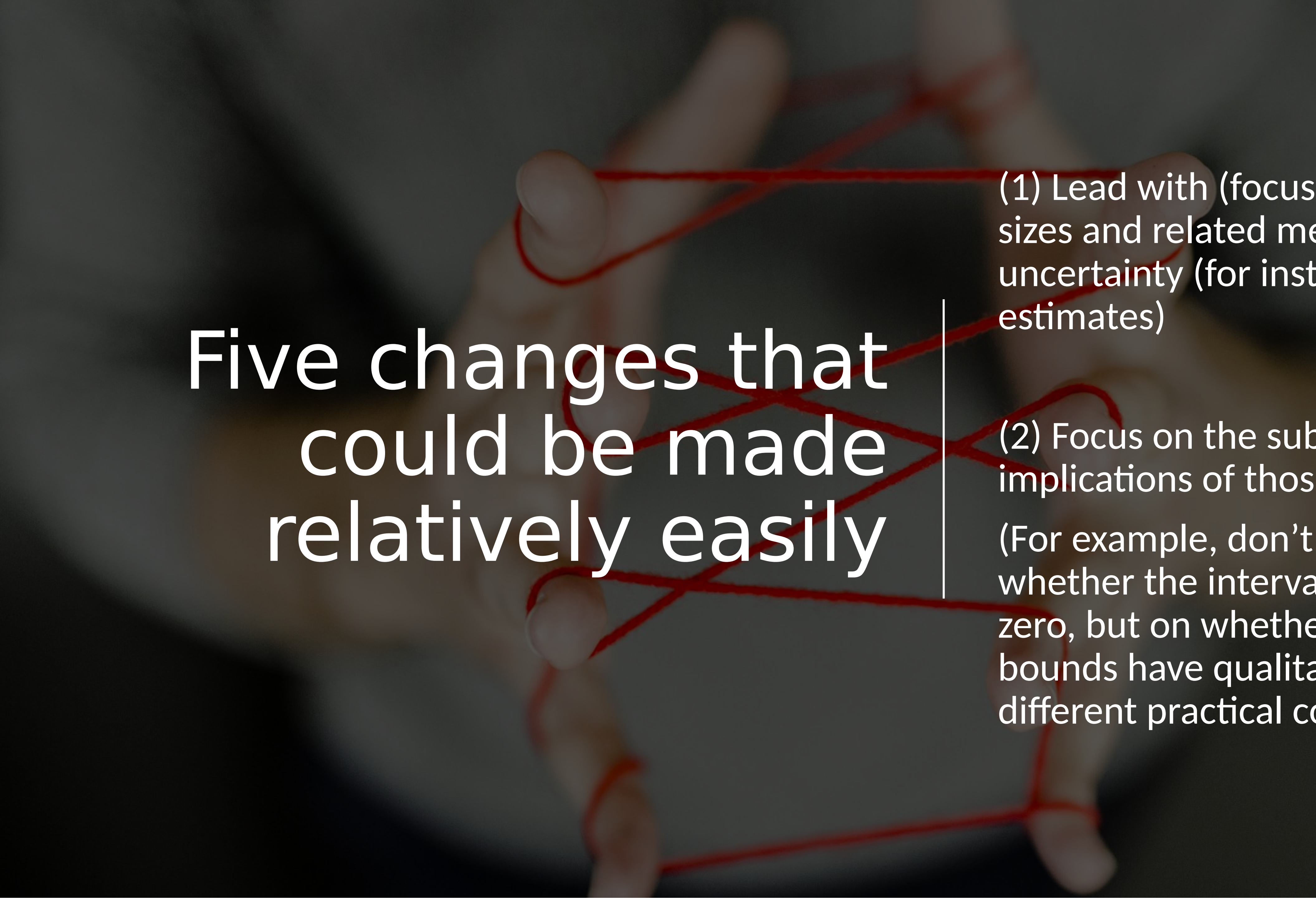
It is reasonab

Do we overst statistics can when we mal arguments?

Image f
eech TV

What do we do instead?

- If we are telling every using thresholds to in what should we do?
- Look for some answe 2019 special issue of *Statistician* (online an
- We'll talk about a few
- As you think about m beyond p<0.05, ask y arbitrary threshold ha created, what would to get your paper pub research grant funded approved, your policy recommendation acc

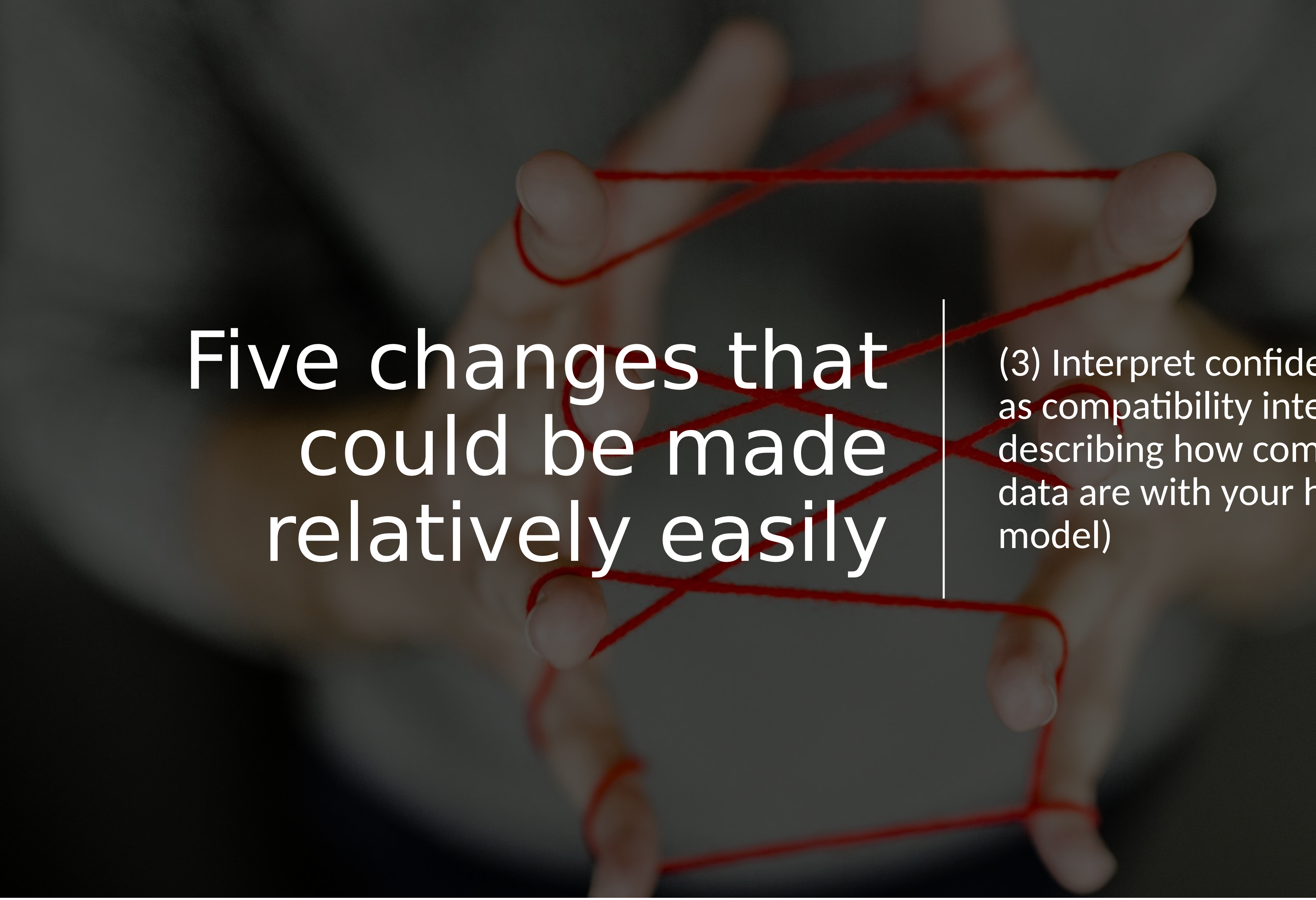


Five changes that could be made relatively easily

(1) Lead with (focus
sizes and related me
uncertainty (for inst
estimates)

(2) Focus on the sub
implications of thos

(For example, don’t
whether the interva
zero, but on whethe
bounds have qualita
different practical c



Five changes that could be made relatively easily

(3) Interpret confide
as compatibility inte
describing how com
data are with your h
model)

Example of compatibility interval interpreta

Study: Covid-19 patients received lopinavir–ritonavir in addition t standard care or standard care alone (randomized trial) (NEJM, M 2020, DOI: 10.1056/NEJMoa2001282)

Result: Mortality difference at 28 days of −5.8 percentage points, (−17.3, 5.7)

Conclusion: “Mortality at 28 days was similar in the lopinavir–rito group and the standard-care group (19.2% vs. 25.0%). … In hospit adult patients with severe Covid-19, no benefit was observed wit lopinavir–ritonavir treatment beyond standard care.”

A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covi

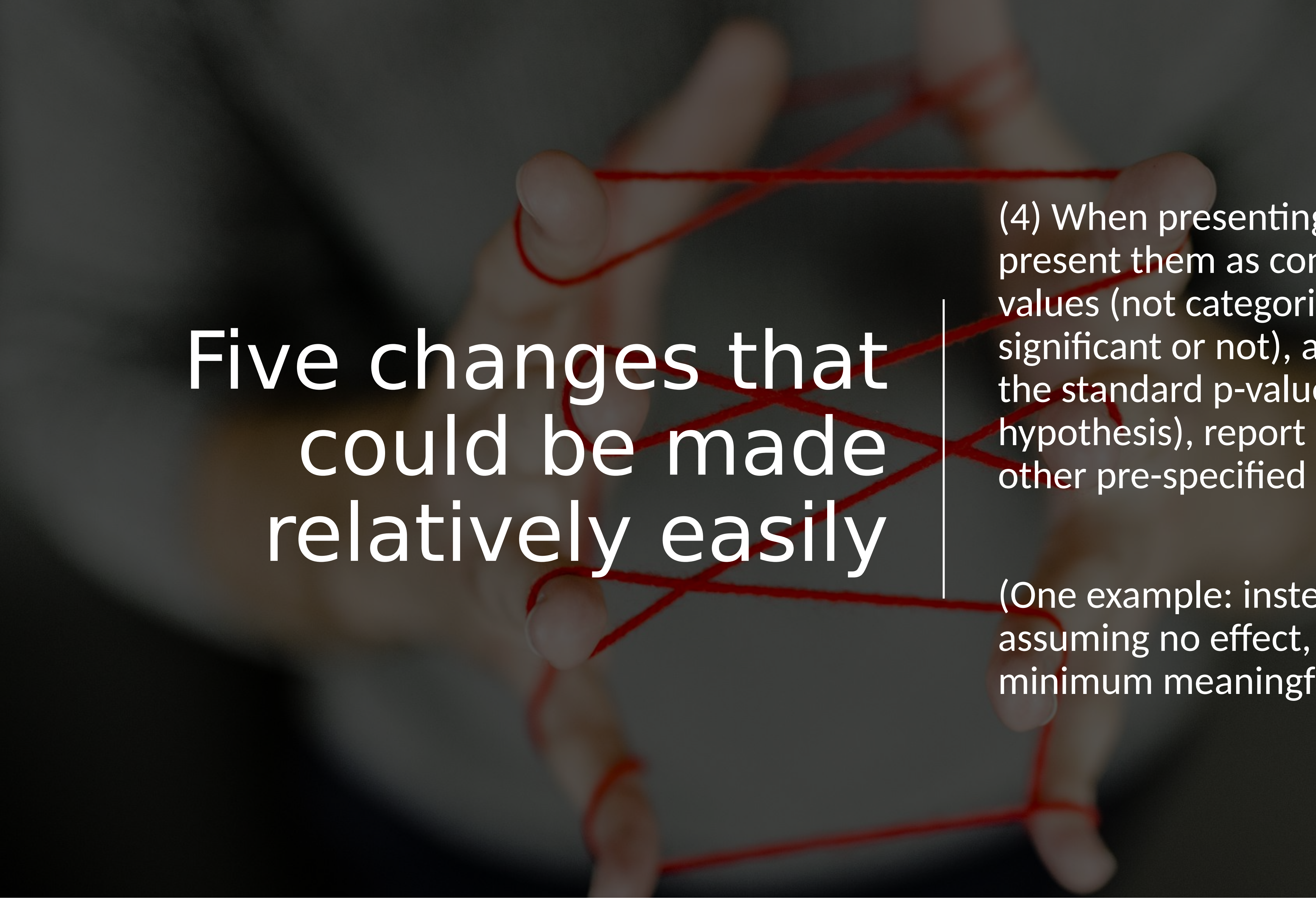
Bin Cao, M.D., Yeming Wang, M.D., Danning Wen, M.D., Wen Liu, M.S., Jingli Wang, M.D., Guohui Fan, M.S., Lianguo Ruan, M.D., Bin Song, M.D., Ming Wei, M.D., Xingwang Li, M.D., Jiaan Xia, M.D., et al.

Example of compatibility interval interpretation

A better statement of this result:

“Our estimate of the mortality difference at 28 days was –5.8 percentage points (= 19.2% – 2
adding lopinavir-ritonavir to standard care could result in a clinically large decrease in mortal
possible mortality differences that are highly compatible with our data, given our model, ran
17.3 (a very large decrease in mortality) to 5.7 (a large increase in mortality). Our trial was sm
only 199 patients, all with severe Covid-19. Further study of this potentially effective treatme

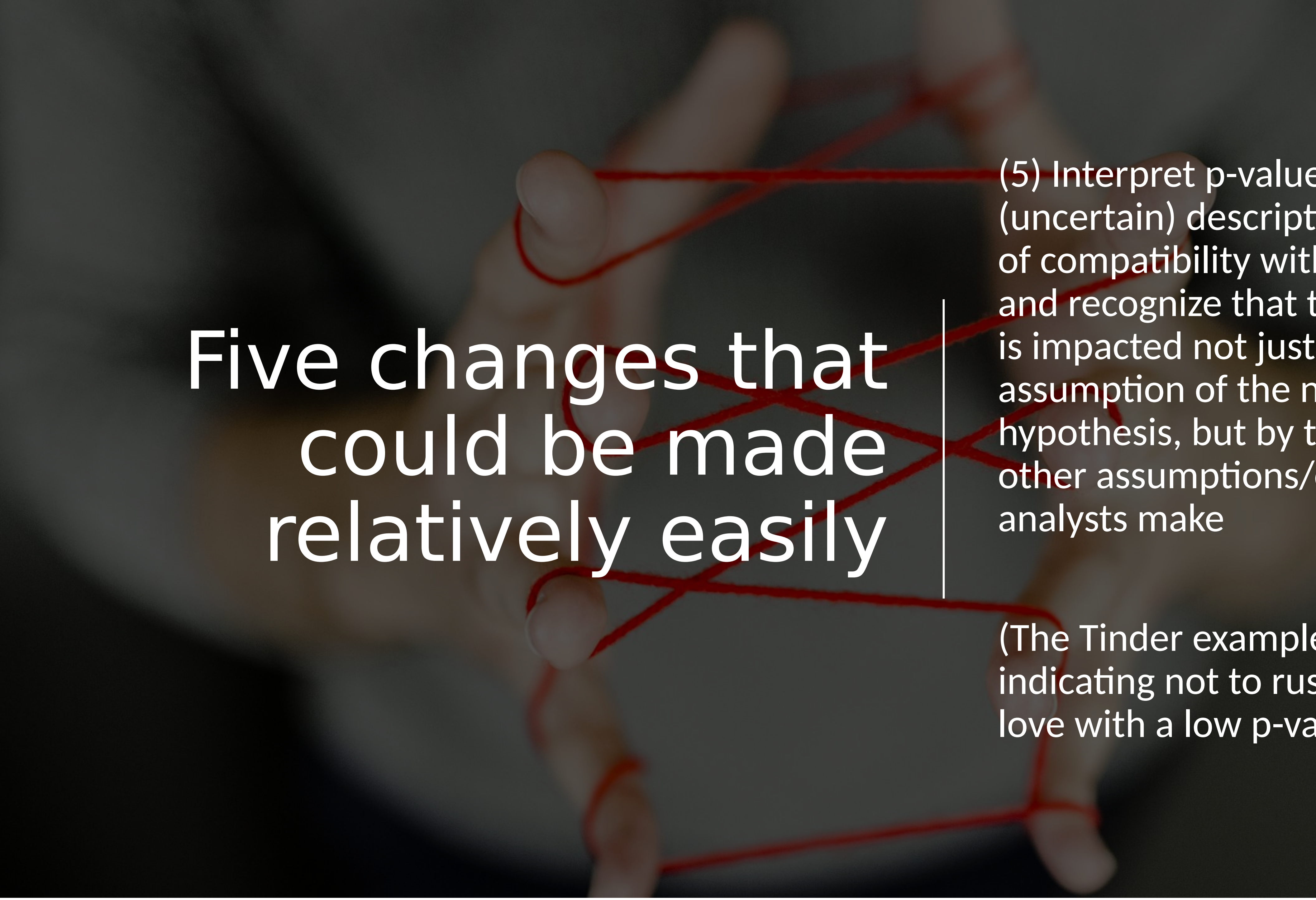
This result should be discussed in the context of the plausibility of the causal mechanism for
effect (based on prior evidence), the high consistency of results across different study outcor
limitations (including but not limited to the large imprecision of the estimates), potential adv
lopinavir-ritonavir, and other relevant considerations.



Five changes that could be made relatively easily

(4) When presentin
present them as co
values (not categori
significant or not), a
the standard p-valu
hypothesis), report
other pre-specified

(One example: inste
assuming no effect,
minimum meaningf



Five changes that could be made relatively easily

(5) Interpret p-value
(uncertain) descript
of compatibility with
and recognize that t
is impacted not just
assumption of the n
hypothesis, but by t
other assumptions/
analysts make

(The Tinder example
indicating not to rus
love with a low p-va

d one more change, a ittle harder, but maybe most important

Don’t focus on the statistical measure alone (for exa the p-value) but also consider

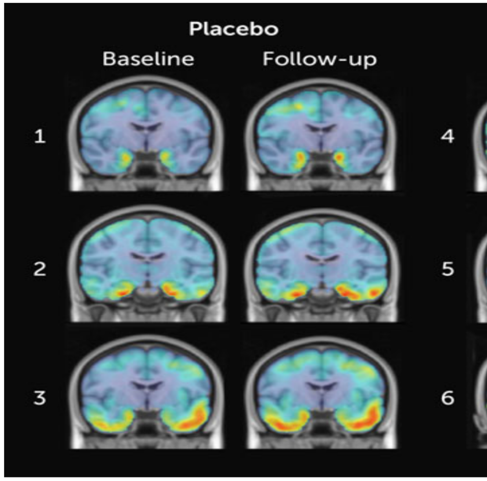
- related prior evidence
- plausibility of mechanism
- study design and data quality
- real world costs and benefits
- novelty of finding
- other factors that vary by research domain

(per McShane et al)

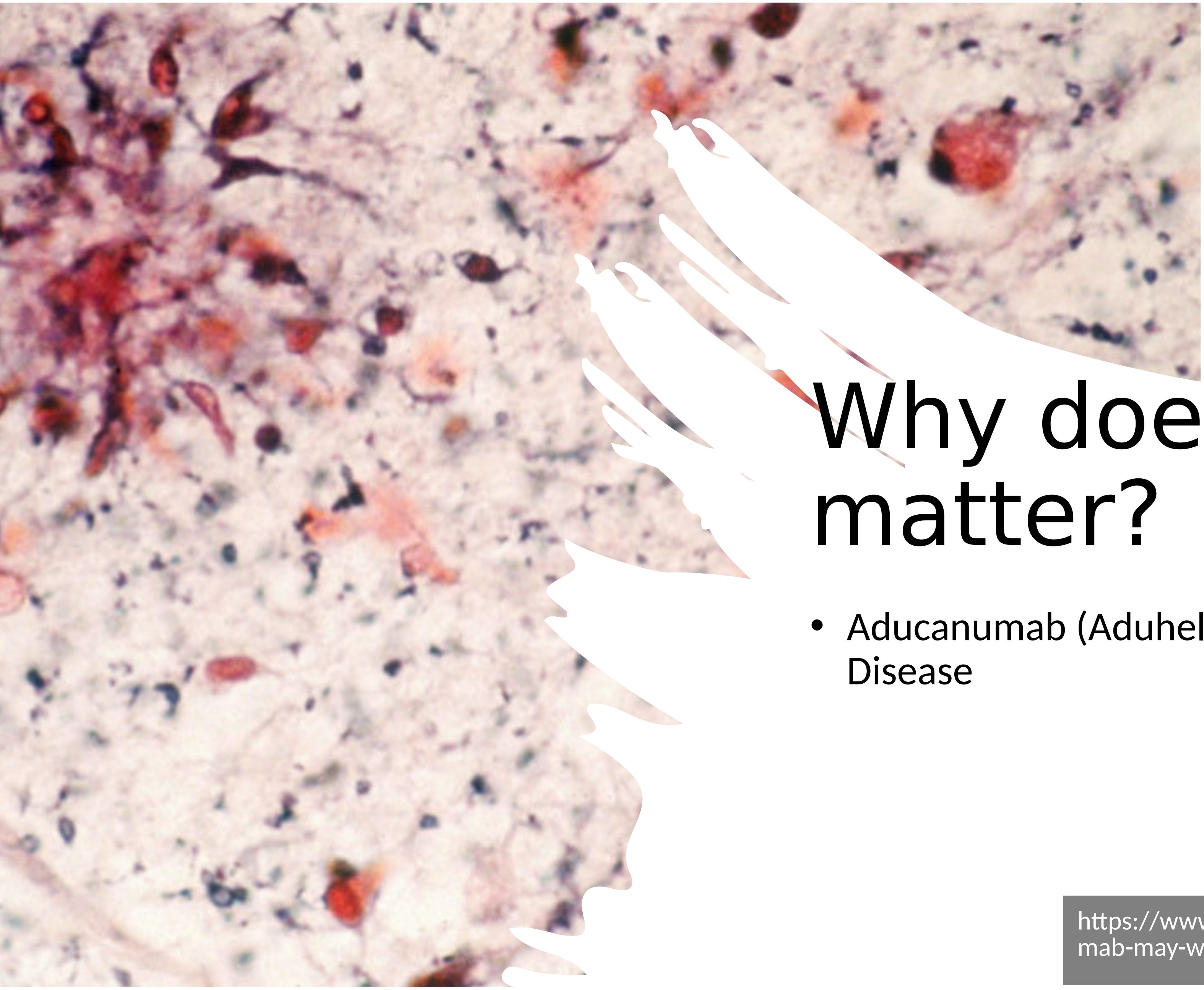
“If this arbitrary threshold h
never been created, what w
you have to do to get your
published, your research gr
funded, your drug approved
policy or business
recommendation accepted

My answer is that you would “have to
the sh*t out of this.” – Mark Watney T
Martian

Progression of tau tangles in individu



BIOGEN



ed) in the brains of people with Alzheimer's, a new drug

Why does this matter?

- Aducanumab (Aduhelm) as a treatment for Alz Disease

https://www.sciencenews.org/article/once-scrapp mab-may-work-after-all

The plot elements

The drug aducanumab, an antibody, has been shown to remove amyloid clusters from the brain.

Such buildup o
associated wit
Disease.

The question is whether removal of amyloids would reduce the effects of Alzheimer’s

No drug has th
succeeded in r
effects

The plot elements

Biogen stopped two simultaneous clinical trials on the effectiveness of aducanumab in March 2019 after futility analysis indicated the study would not likely demonstrate efficacy.

However, more data c

“Between December 2018, when data were cut for the futility analysis, and March 2019, when the trials were discontinued, an additional 179 EMERGE and 139 ENGAGE participants completed 18 months of follow-up”

Howard, R., and Liu, K.
“Questions EMERGE as
claims aducanumab tu
Nature Reviews Neuro
https://doi.org/10.103
019-0295-9.

The plot elements

A subset analysis was undertaken of those participants who re
the full, uninterrupted treatment

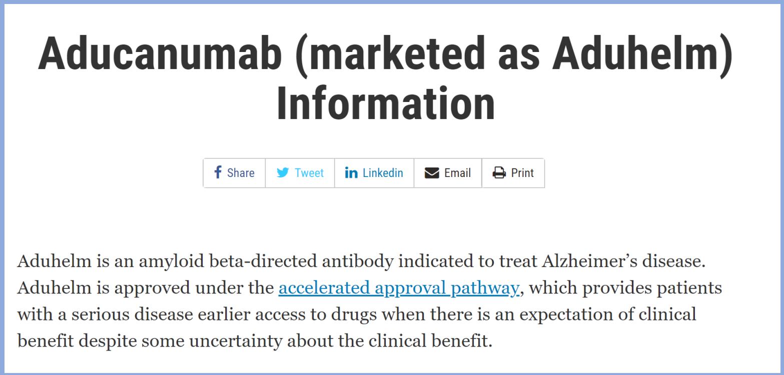
In ONE of the two trials, statistical significance was achieved.
higher dose led to 23% less cognitive decline than a placebo a
weeks.

The plot thickens

Biogen argues that the difference in the results can be explained by a protocol change, but this is based on p subgroup analysis, not the best place to focus on p-va

The effect sizes may not actually meet a threshold of clinical significance.

What happened? FDA approval (June 2021)

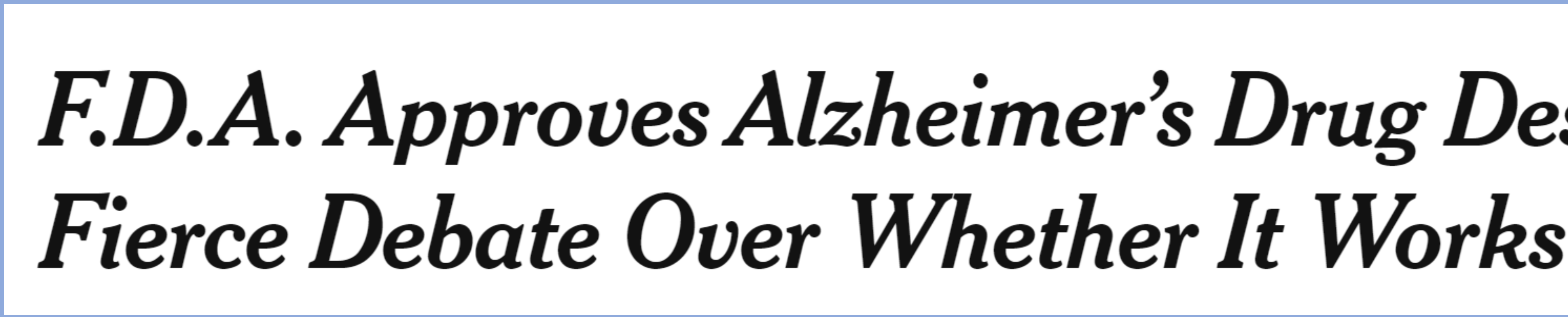


Aducanumab (marketed as Aduhelm) Information

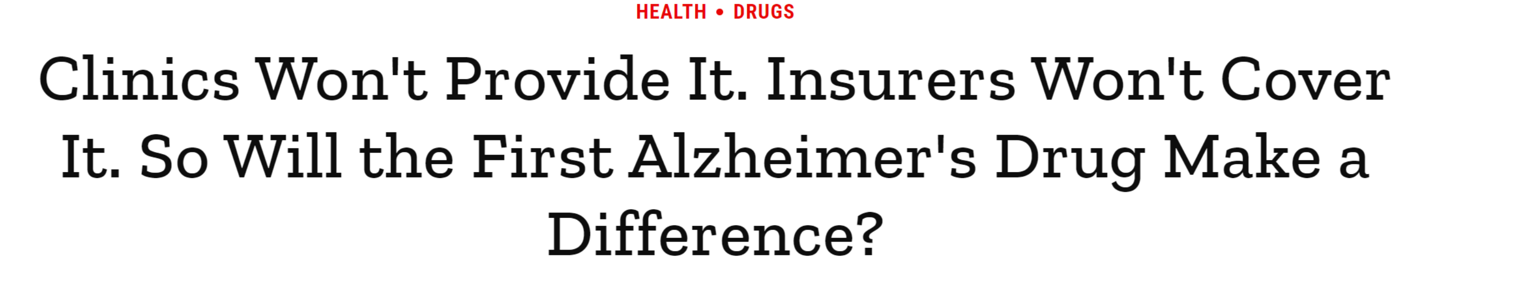
Aduhelm is an amyloid beta-directed antibody indicated to treat Alzheimer's disease. Aduhelm is approved under the accelerated approval pathway, which provides patients with a serious disease earlier access to drugs when there is an expectation of clinical benefit despite some uncertainty about the clinical benefit.

FDA U.S. FOOD & DRU

ADMINISTRATION

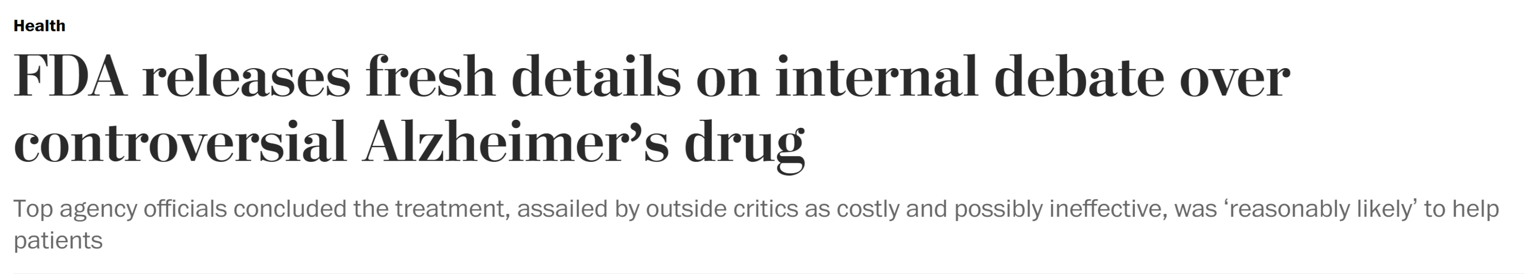


F.D.A. Approves Alzheimer's Drug De
Fierce Debate Over Whether It Works



HEALTH • DRUGS

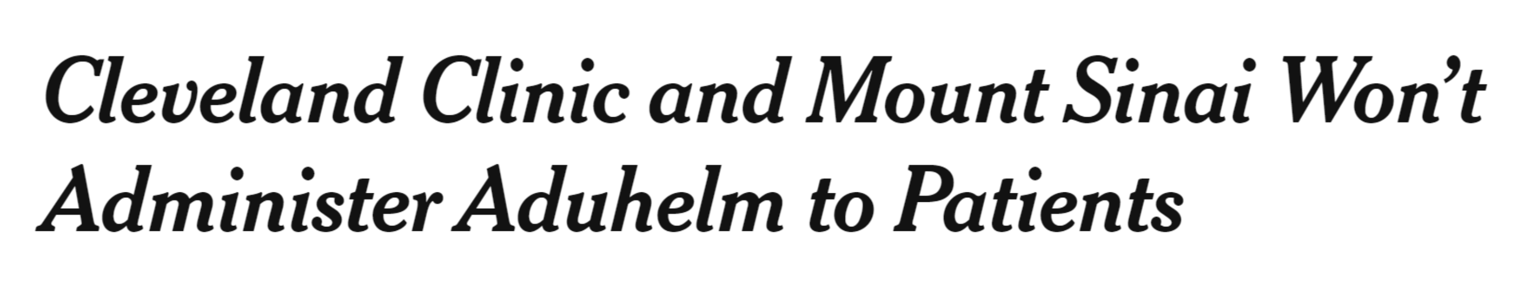
Clinics Won't Provide It. Insurers Won't Cover It. So Will the First Alzheimer's Drug Make a Difference?



Health

FDA releases fresh details on internal debate over controversial Alzheimer's drug

Top agency officials concluded the treatment, assailed by outside critics as costly and possibly ineffective, was 'reasonably likely' to help patients



Cleveland Clinic and Mount Sinai Won't Administer Aduhelm to Patients

But a
upro
arose

Statistical significance?

We’re not privy to all the internal workings

We aren’t experts (but the internal FDA committee members ARE)

Impact of focusing on a threshold – apparent p-hacking kept the product a

Lots of money and hopes involved

Lecanemab (Leqembi) was approved in January 2023

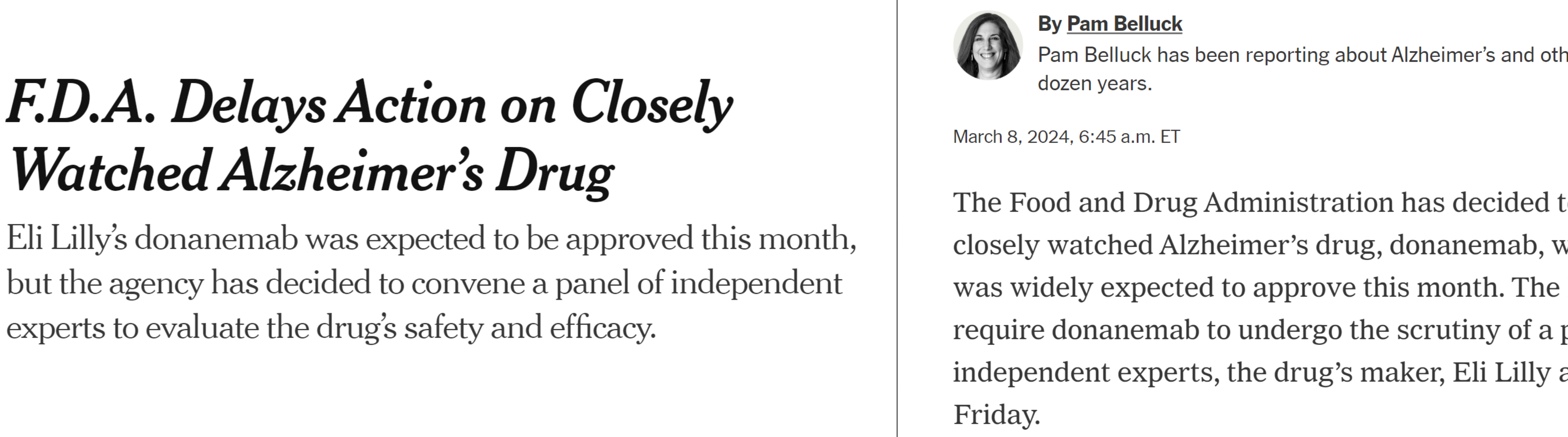
“Still, several Alzheimer’s experts said it was unclear from the medical evic whether Leqembi could slow cognitive decline enough to be noticeable to patients.” FDA Approves, Leqembi, New Treatment for Early Alzheimer’s - The New York Times (nytimes.co

But then…

January 31, 2024 – Biogen stops tests and abandons the drug

https://www.nytimes.com/2024/01/31/business/biogen-alzheimers-aduh

And then March 7, 2024



F.D.A. Delays Action on Closely Watched Alzheimer's Drug

Eli Lilly's donanemab was expected to be approved this month, but the agency has decided to convene a panel of independent experts to evaluate the drug's safety and efficacy.

By Pam Belluck
Pam Belluck has been reporting about Alzheimer's and oth
dozen years.

March 8, 2024, 6:45 a.m. ET

The Food and Drug Administration has decided t
closely watched Alzheimer's drug, donanemab, w
was widely expected to approve this month. The
require donanemab to undergo the scrutiny of a p
independent experts, the drug's maker, Eli Lilly a
Friday.

https://www.nytimes.com/2024/03/08/
heimers-drug-donanemab.html

Wrapping up

- It’s time to stop using “st
significance” as any kind
metric for scientific infere
and teaching it as a found
concept
- We and many others hav
written a lot about what
beyond P<0.05” should lo
- P-values still have their u

“(S)cientists have embraced and even avidly **pursued meaning 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 ha suffered or died** because scientists (and editors, regulators, journalists and others) have used significance tests to interpre results, and have consequently failed to identify the most beneficial courses of action.”

- (Rothman, supplement to the 2016 ASA statement)

Thanks for your time and attent

Please send comments to ron@amstat.org.

And remember the disclaimer!