Tackling dichotomania (p<0.05) and other statistical diseases in pursuit of better science

John Carlin

ACTA-STInG 11-Aug-2026

murdoch children's research institute | CEBU Clinical Epidemiology and Biostatistics Unit | The University of Melbourne

1

Outline

- Dichotomania: 10 years since the ASA statement and Greenland et al…
- Statistical significance: what exactly is the problem?
- Progress on solving the problem?
- Potential strategies connecting to "other statistical diseases"

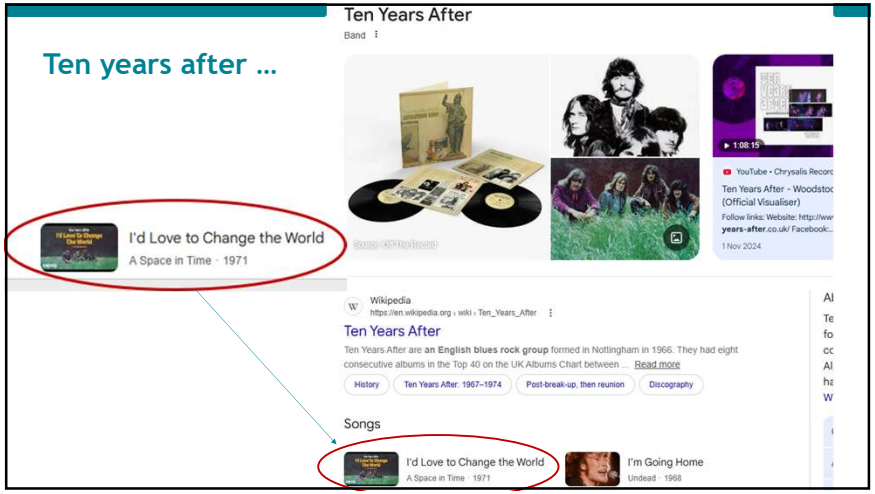
2

Ten years after …

… the American Statistical Association released a statement recommending against the dichotomous interpretation of p-values, what has changed and what still needs to change?

3

Ten years after …

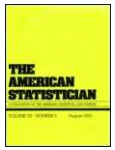


4

Ten years since…

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, DOI: 10.1080/00031305.2016.1154108

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

5

Eight years since a landmark paper…

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

CrossMark

ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

- **4,357 citations!** (Altmetric 28/07/26)

6

Seven years since…

nature

Explore content · About the journal · Publish with us · Subscribe

nature › comment › article

https://www.nature.com/articles/d41586-019-00874-8

COMMENT | 20 March 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenla
for an end to hyped claims and the

By Valentin Amrhein, Sander Greenlan

MENU nature International journal of science

Subscribe

EDITORIAL · 20 MARCH 2019

It's time to talk about ditching statistical significance

Looking beyond a much used and abused measure would make science harder, but better.

7

Six years since my last talk on this topic…

Beyond dichotomous thinking in clinical trials: why, how & who?

John Carlin

Murdoch Childrens Research Institute
& University of Melbourne

ISCB 2020
26-Aug-20

Krakow, Poland

[or the front room at home in Melbourne at 2am (COVID!!)]

8

With a literature going back decades…

Ziliak ST, McCloskey DN. The cult of statistical significance: how the standard error costs us jobs, justice and lives. Ann Arbor: U Michigan Press; 2008.

Goodman SN. Towards evidence-based medical statistics, I: the P-value fallacy. *Ann Intern Med*. 1999;130:995–1004.

Chia KS. "Significant-itis"—an obsession with the P-value. *Scand J Work Environ Health*. 1997;23:152–4.

Cohen J. The earth is round (p<0.05). *Am Psychol*.1994;47:997–1003.

Gardner MA, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J*. 1986;292:746–50.

Morrison DE, Henkel RE, editors. *The significance test controversy*. Chicago: Aldine; 1970.

Rozeboom WM. The fallacy of null-hypothesis significance test. *Psychol Bull*. 1960; 57:416–28.

Berkson J. Tests of significance considered as evidence. *J Am Stat Assoc*. 1942;37:325–35.

Boring, EG. Mathematical vs. scientific significance. *Psychol Bull*. 1919; 16(10):335–338.

9

9

So, what exactly is the dichotomania (a/k/a significantitis) problem?

10

10

An example from the *New England Journal of Medicine*

N ENGL J MED 371;18 NEJM.ORG OCTOBER 30, 2014

ORIGINAL ARTICLE

Simvastatin in the Acute Respiratory Distress Syndrome

Results:

"There was no significant difference between the study groups in the mean (±SD) number of ventilator-free days (12.6±9.9 with simvastatin and 11.5±10.4 with placebo, P = 0.21)…"

Conclusions:

"Simvastatin therapy, although safe and associated with minimal adverse effects, did not improve clinical outcomes…"

11

11

"… did not improve clinical outcomes…"

Null hypothesis not rejected, leading to conclusion of "no effect"…

BUT WAIT:

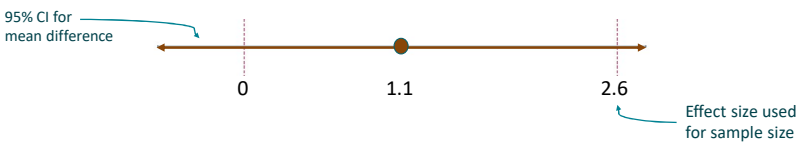
"Sample-size assumptions […] a sample of 524 patients […] in order for the study to have 80% power, at a two-tailed significance level of 0.05, to detect a mean between-group difference of 2.6 ventilator-free days…"

12

12

"… did not improve clinical outcomes…"

Let's look at the results again:



95% CI for mean difference: -0.6 to 2.8

The "alternative hypothesis" should not be rejected either!!!

13

13

"… did not improve clinical outcomes…"

What should the reader take away from this paper?

- NOT: "the drug improves clinical outcomes" (so we should use it?)
- NOT: "the drug has no effect on clinical outcomes" (so we should forget about it?)

Clearly a false dichotomy!

- In fact, there was uncertain evidence for potential benefit

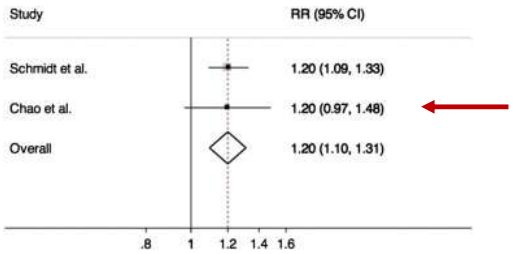
Consequences of dichotomising

- Confusion about what the trial results mean, because stakeholders expect something that can't be provided (certainty?!)

14

14

Another (extreme!) example



"A unique finding of the present study was that the use of selective COX2 inhibitors was not associated with AF risk, except in patients with chronic renal or pulmonary disease."

FIG. 2 Meta-analysis of the association between use of COX-2 inhibitors and atrial fibrillation as reported by Schmidt et al. and Chao et al.

Schmidt & Rothman (2014) "Mistaken inference caused by reliance on and misinterpretation of a significance test" *Int.J.Cardiol.*

15

15

Anonymous correspondence with *JAMA* (2020)

Authors' submitted manuscript reported primary result as:
RR 1.05; 95% confidence interval 1.00 to 1.10

Editor's response:

For the primary outcome, the P-value is presumably slightly above .05 […] To help make this clear, please carry the P-value, for the primary outcome only, out to 3 decimal places.

Final version:
RR 1.05; 95% confidence interval 0.999 to 1.098, p = 0.054

At editor's insistence, this was reported as "did not meet statistical significance"

16

16

Anonymous correspondence with *Lancet* (2017)

Copy Editor rounded results to one digit:

However, they were less likely to be admitted to hospital with depressive mood disorder (IRR 0·7, 95% CI 0·7–0·8) …

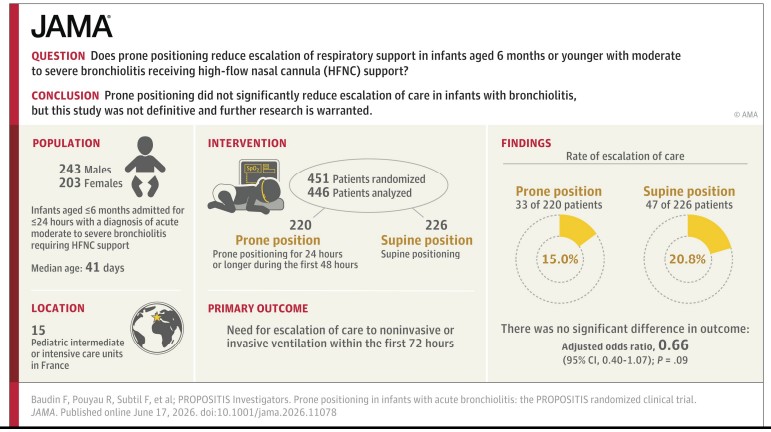
Author requested: "please restore second decimal place…", editor responds:

A: thank you for your suggestions, I am happy to include these data to 2 d.p.; however, I would like to point out that the addition to 2 d.p., strictly speaking, would not change the statistical significance of the reported data, as such these changes so late in the publication process begs the question of whether such accuracy is truly necessary.

17

17

An example from latest issue of *JAMA*



18

18

Dichotomania: why is it a problem?

Claims of "significant"

- It is not clear what should follow from the claim (change of practice, adoption of intervention etc), but typically the only clear result is a "ticket" for publication i.e. for claims of importance to science or medicine, etc.
  - Selective reporting is clearly incentivised
  - Selective reporting of "significant" findings leads to publication of inflated estimates of effects

Claims of "non-significant"

- Common conflation of "not significant" with "no effect" is clearly wrong and a source of much confusion and potential downplaying of potentially important effects (both benefit and harm)

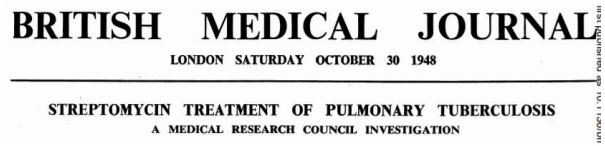
What do such claims even mean (logically), especially given that the threshold for "significance" (traditionally 0.05) is arbitrary? Will return to this…

19

19

Why the dichotomy?

- It goes back a long way… e.g. this famous (first in medicine?) randomised trial:


BRITISH MEDICAL JOURNAL

LONDON SATURDAY OCTOBER 30 1948

STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS

A MEDICAL RESEARCH COUNCIL INVESTIGATION

"Four of the 55 S patients (7%) and 14 of the 52 C patients (27%) died before the end of six months. **The difference between the two series is statistically significant**; the probability of it occurring by chance is less than one in a hundred."

N.B. This is one of the classic misinterpretations!

20

20

Why the dichotomy?

Complex historical/sociological origins in science broadly

- Strong (instinctive/intuitive?) appeal of a yes/no answer
  - Shouldn't statisticians know better?
- Appeal of "objectivity," tied to quantification of knowledge
  - Gigerenzer G, Marewski JN. Surrogate science: the idol of a universal method for scientific inference. *J Management*. 2015

In trials more specifically?

- Linked to regulatory decision-making
  - Drug can proceed to next stage of approval if trial "succeeds"
- … but how many trials lead directly to an actual "decision"?
  - And if a decision, how often is that driven purely by the statistical significance attached to the primary outcome comparison?

21

21

Looking more closely at the dichotomy

Early 20th century: Statistical inference became conflated with drawing a dichotomous conclusion…

- Neyman-Pearson theory of hypothesis testing appeared to fit the bill

BUT Neyman & Pearson (1933) explicitly stated that their theory of testing was not designed for drawing inferences from specific datasets!

"Without hoping to know whether each separate hypothesis is true or false, we may search for ***rules to govern our behaviour*** **with regard to them, in following which we ensure that, in the long run of experience, we shall not often be wrong**."

In contrast, Fisher (1925) emphasised the *P*-value:

- "Index of surprise": if small, "consider alternative explanations…"
- This was inference *from the data*…

22

22

Looking more closely at the dichotomy

Standard practice of null hypothesis significance testing mashes up Neyman-Pearson theory with Fisher and has no logical basis!

- Define a test statistic T and calculate $P = \Pr(T > t_{obs} \mid H_0)$
- If $P < 0.05$ then declare that 'statistical significance' has been observed, implying that H_0 has been rejected/disproven
- This is claiming a scientific inference from the *P*-value, i.e. drawing a conclusion *given* the data (not just engaging in "good long-run behaviour")

IT DOESN'T MAKE SENSE!!

- Confusion about Type I/II error rates cf. actual "decision"
- Misinterpretation in both directions ("significant" and "non-significant")
  - Currency of publication, Type M error (winner's curse), *P*-hacking, replicability

Goodman S (1999) "Toward evidence-based medical statistics. 1: The P value fallacy" *Ann Int Med*

23

23

Are trials different? Is there a special case for hypothesis tests in trials?

Stricter rules (prespecified primary comparison at prespecified alpha level, etc) enforce discipline re primary analysis, power, multiple comparisons etc

- But still fundamentally illogical and misleading
  - Trial reports should not confuse the reporting of estimated effects with conclusions or decisions about recommended practices

Isn't the machinery of Type I and Type II error control critical to maintaining trial quality?

- No! Individual trials never "determine" conclusions about whether a treatment effect is or is not zero
- This is a game that is maintained in order to provide a framework for determining sample size
  - Increasingly in question as Bayesian analysis methods proliferate?

24

24

Looking at the dichotomy: summary

***P* < .05 does NOT mean an effect is real**

***P* > .05 does NOT mean there is no effect**

(And numerous other common misunderstandings!)

- Statisticians should not perpetuate the practice of drawing (falsely) dichotomous conclusions

25

25

Have things changed (longer term and recently)?

- **Practice of major journals**
- **Trends more broadly in medical journals**

26

26

Major journals set the tone: have they changed? e.g. *JAMA*

- **From guidelines to authors, "Statistical Methods and Data Presentation"**

Avoid solely reporting the results of statistical hypothesis testing, such as *P* values, which fail to convey important quantitative information. For most studies, *P* values should follow the reporting of comparisons of absolute numbers or rates and measures of uncertainty (eg, 0.8%, 95% CI −0.2% to 1.8%; *P* = .13). *P* values should never be presented alone without the data that are being compared.

- OK, but what does "comparison of absolute numbers or rates" mean?! (etc.)

27

27

Major journals set the tone: have they changed? e.g. *JAMA*

[guidelines cont.]

- **"significance" mentioned four times:**

Do not use inappropriate hedge terms such as marginal significance or trend toward significance for results that are not statistically significant.

State any a priori levels of significance and whether hypothesis tests were 1- or 2-sided.

For secondary and subgroup analyses, there should be a description of how the potential for type I error due to multiple comparisons was handled, for example, by adjustment of the significance threshold.

- Overall message is confusing!

28

28

Major journals: *NEJM*

https://www.nejm.org/author-center/new-manuscripts#statistical-reporting-guidelines

Under "1. *General requirements and reporting style*"

c. To prevent confusion, we recommend that authors draw a clear distinction between statistical significance and clinical (or other nonstatistical) significance. This can be achieved by reserving the adjective "significant" to mean "statistically significant."

d. Significance tests should be accompanied by effect estimates with standard errors or confidence intervals.

OK, perhaps?? (assuming that significance declarations are meaningful!)

29

29

Major journals: *NEJM*

Under "2. *Multiplicity considerations*" they **dig a big hole**!

a. To enhance the reproducibility of study results, the *Journal* emphasizes the need to account for the multiplicity of hypothesis testing in the analysis and to control the probability of error. […]

b. For confirmatory analyses, such as those typically reported for end points in clinical trials, the SAP should prespecify a plan to control overall type I error (family-wise error rate). […]

d. When no method to adjust for multiplicity of inferences was prespecified in the protocol or SAP, reporting of secondary and exploratory end points should be limited to point estimates of effects with 95% confidence intervals. In such cases, the Methods section should state that the widths of the intervals have not been adjusted for multiplicity and that the intervals may not be used in place of hypothesis testing. The interpretation of these confidence intervals should avoid the language of definitive conclusions used to report statistically significant findings as assessed by formal hypothesis testing.

30

30

Major journals: *BMJ*

Under "Statistical issues":

"Please do not use the term 'negative' to describe studies that have not found statistically significant differences, perhaps because they were too small. There will always be some uncertainty, and we hope you will be as explicit as possible in reporting what you have found in your study. Using wording such as "our results are compatible with a decrease of this much or an increase of this much" or 'this study found no effect' is more accurate and helpful to readers than "there was no effect/no difference." "

Muddled!!

- Overall, no clear progress in top journals ☹
- What about bigger picture?

31

31

Quantitative trends in practice

Eur J Epidemiol (2017) 32:21–29
DOI 10.1007/s10654-016-0211-1

REVIEW

Statistical inference in abstracts of major medical and epidemiology journals 1975–2014: a systematic review

Andreas Stang[1,2] · Markus Deckert[1] · Charles Poole[3] · Kenneth J. Rothman[4]

- 5 major medical journals (*NEJM*, *JAMA*, *Lancet*, *BMJ*, *AnnIntMed*)
  7 epidemiology journals (*AmJEpi*, *IJE*, *Epid*, *EurJEpi*, *JECH*, *JCE*, *AnnEpi*)
- "CI-only" approach rose monotonically over time in all journals
- BUT "NHST" retained dominance in 4/5 medical journals though not epi journals
- "The reporting of CIs without explicitly interpreting them as statistical tests is becoming more common in abstracts, particularly in epidemiology journals. Although NHST is becoming less popular in abstracts of most epidemiology journals studied and some widely read medical journals, it is still very common in the abstracts of other widely read medical journals, especially in the hybrid form of ST and NHT in which p values are reported numerically along with declarations of the presence or absence of statistical significance."

32

32

Evolution of Reporting P-values Across the Biomedical Literature, 1990-2025: an Updated Meta-Research Study

Posted January 16, 2026. medRxiv

Jinhyeok Choi, Keeheon Lee, David Chavalarias, Jae Il Shin, John P A Ioannidis

doi: https://doi.org/10.64898/2026.01.14.26344149

Main Outcomes and Measures Proportion of article reporting at least 1 *P*-value, at least 1 *P*-value less than thresholds (.05 and .005), distribution of *P*-values by magnitude and operator type.

Results The proportion of articles reporting *P*-values increased from 7.5% in 1990 to 18.3% in 2025 for PubMed abstracts, and from 5.2% to 53.3% for PMC full-texts. The median number of *P*-values per article increased from 2 to 7 in PMC full-text articles, with upward trends observed in all databases. A high proportion of *P*-values remains clustered around .05 and .001 in all databases. The proportion of articles reporting at least one *P*-value ≤ .05 has remained in the range 94%-98% since 1998, while the proportion reporting at least one *P*-values ≤ .005 has increased over time, reaching 57.0% for PubMed abstracts and 62.5% for PMC full-texts. The reporting of 'exact' *P*-values increased until 2015, but with no further increase in the last 10 years (PubMed abstracts: 17.6% in 1990, 51.1% in 2015, 49.8% in 2025)

Conclusions and Relevance Our evaluation demonstrates the pervasive entrenchment of *P*-values, despite heavy debates and major changes in the content of the biomedical literature over time. More *P*-values are reported and papers using *P*-values almost always report some that are statistically significant. Readers should remain aware of the major issues surrounding *P*-value misuse and misinterpretation.

33

33

European Journal of Epidemiology (2023) 38:1035–1042
https://doi.org/10.1007/s10654-023-01047-8

REVIEW

Statistical inference and effect measures in abstracts of randomized controlled trials, 1975–2021. A systematic review

Andreas Stang[1,2] · Kenneth J Rothman[2]

"Reporting of confidence intervals increased from 0.5% (1975) to 29% (2021)."

BUT:

"Overall and within each calendar time period, significance language was the most common mode of statistical inference. Even in the most recent period, 2015-2021, significance language was used in 77% of all abstracts."

[And, the most common effect measures were hazard ratio and odds ratio...]

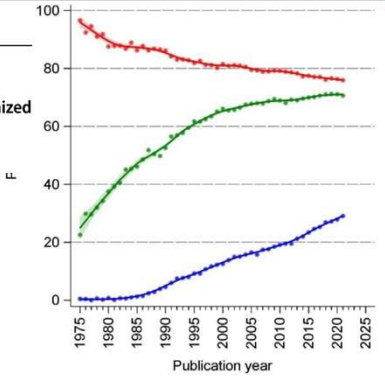


Fig. 2 Estimated time trends, 1975–2021, of the statistical reporting style in abstracts on randomized controlled trials (non-hierarchical) **Blue** indicates reporting of confidence intervals; **green** indicates reporting of numerical p-values or p thresholds, (e.g. $p<0.01$); **red** indicates reporting of significance language. All trend lines are LOESS smoothed with inverse-variance weighting

34

ELSEVIER journal homepage: www.sciencedirect.com/journal/global-epidemiology

Review Article

Statistical inference and effect measures in abstracts of major HIV and AIDS journals, 1987–2022: A systematic review

Andreas Stang [a,b,c,*], Henning Schäfer [d,e], Ahmad Idrissi-Yaghir [d,f], Christoph M. Friedrich [a,d], Matthew P. Fox [c]

- Latest in a series led by Andreas Stang looking at various areas of medical research (cancer, clinical pharmacology, cardiovascular)
- This one showed the strongest evidence for changes in practice
- Clear increase in reporting of CIs (although noting this may not be free of dichotomania, and also querying choice of effect measures)

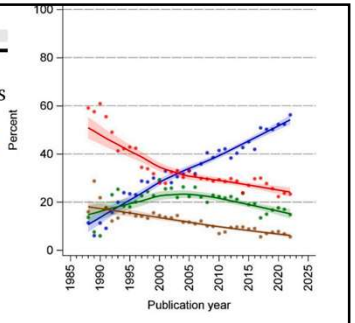


Fig. 2. Annual reporting style on statistical inferences in abstracts that contain any statistical inference in 10 major HIV/AIDS journals from 1987 through 2022.
Legend: **blue** indicates reporting of confidence intervals; **green** indicates reporting of numerical *p*-values (e.g. $p = 0.03$) without confidence intervals; **brown** indicates reporting of *p*-value thresholds (e.g. $p < 0.01$) without confidence intervals and numerical *p*-values; **red** indicates reporting of significance without confidence intervals, numerical p-values and *p*-value thresholds; all trend lines are LOESS smoothed with inverse-variance weighting; the colored bands present the 95 % confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

35

Trends to (away from?) significance…

European Journal of Pediatrics (2023) 182:937–940
https://doi.org/10.1007/s00431-022-04746-8

BRIEF REPORT

The trends for the "trend toward significance" in the pediatric literature

AOA Critical Issues in Education

The Trend to Use the Word "Trend" to Describe Nonsignificant Results in Orthopaedic Literature

Patrick A. Massey, MD, MBA, FAOA, Ryan Taylor, MD, Robert W. Rutz, MD, Joshua Harris, MD, Carver Montgomery, MD, Mohammad Alfrad Nobel Bhuiyan, PhD

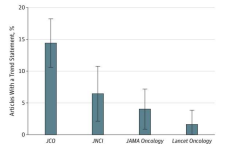
British Journal of Anaesthesia 115 (3): 337–9 (2015) JNCI Journal of the National Cancer Institute.
Advance Access publication 29 May 2015 · doi:10.1093/bja/aev149

RESEARCH LETTER

The Use of "Trend" Statements to Describe Statistically Nonsignificant Results in the Oncology Literature

The use and interpretation of *P* values in the biomedical literature is problematic. The importance of a study is often inappropriately defined by the *P* value. This problem is highlighted by the use of "trend" to refer to statistically nonsignificant results. There is no definition of a trend toward statistical sig-

Figure. Proportion of Original Research Articles Using Trend Statements to Describe 1 or More Statistically Nonsignificant Results From November 2016 Through October 2017



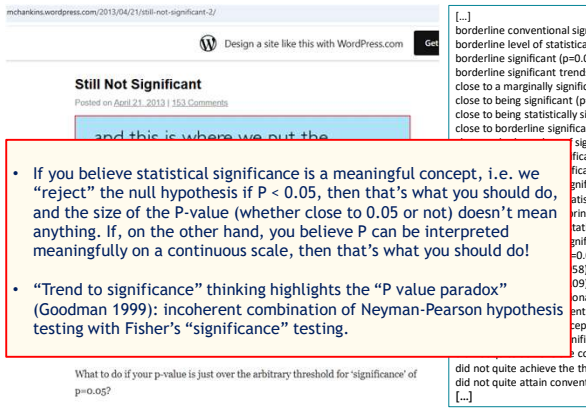
The error bars indicate 95% CIs. JCO indicates Journal of Clinical Oncology, and JNCI, Journal of the National Cancer Institute.

Misuse of 'trend' to describe 'almost significant' differences in anaesthesia research

N. M. Gibbs[1,*] and S. V. Gibbs[2]
[1] Department of Anaesthesia, Sir Charles Gairdner Hospital, Nedlands, WA 6009, Australia, and
[2] The University of Western Australia, Crawley, WA 6009, Australia

36

"Trending to significance" highlights the nonsense of significance



- If you believe statistical significance is a meaningful concept, i.e. we "reject" the null hypothesis if P < 0.05, then that's what you should do, and the size of the P-value (whether close to 0.05 or not) doesn't mean anything. If, on the other hand, you believe P can be interpreted meaningfully on a continuous scale, then that's what you should do!
- "Trend to significance" thinking highlights the "P value paradox" (Goodman 1999): incoherent combination of Neyman-Pearson hypothesis testing with Fisher's "significance" testing.

[…]
borderline conventional significance (p=0.051)
borderline level of statistical significance (p=0.053)
borderline significant (p=0.09)
borderline significant trends (p=0.099)
close to a marginally significant level (p=0.06)
close to being significant (p=0.06)
close to being statistically significant (p=0.055)
close to borderline significance (p=0.072)
significance (p=0.06)
ficance (p=0.07)
ficance (p=0.17)
gnificance (p=0.055)
atistical significance (p=0.075)
rink of significance (p=0.07)
tatistical significance (p=0.0669)
gnificance (p>0.05)
=0.06)
58)
09)
onal threshold levels of statistical significance (p=0.08)
entional level of statistical significance (p<0.08)
ceptable levels of statistical significance (p=0.054)
nificance (p=0.076)
e conventional levels of significance (p=0.052)
did not quite achieve the threshold for statistical significance (p=0.08)
did not quite attain conventional levels of significance (p=0.07)
[…]

37

JNCI Cancer Spectrum, 2024, 8(3), pkae035
https://doi.org/10.1093/jncics/pkae035
Advance Access Publication Date: April 29, 2024
Commentary

The statistical significance revolution

Alfred I. Neugut, MD, PhD[1,2,*], Tito Fojo, MD, PhD[1,3]
[1]Department of Medicine and Herbert Irving Comprehensive Cancer Center, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA
[2]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA
[3]James J. Peters Department of Veterans Affairs Medical Center, Bronx, NY, USA
*Correspondence to: Alfred I. Neugut, MD, PhD, Division of Oncology, Columbia University Irving Medical Center, 722 W 168th St, Room 725, New York, NY 10032, USA (e-mail: ain1@cumc.columbia.edu).

Abstract

Statistical significance has long relied on the criterion of P less than or equal to .05. Although this threshold has generally functioned well, it has engendered some negative practices to circumvent it and been criticized as too inflexible. We concur with the statisticians and methodologists who are currently arguing for more flexibility to the P value and more reliance on the 95% confidence interval, a shift that is likely to change future practice in data analysis and interpretation for oncology.

- This commentary reflects difficulty of issues (persisting confusion around "hypotheses" and "decisions", etc)
- BUT may be indicative of changes in the air?
- … although empirical data are (so far) less reassuring [as above]

38

Tackling dichotomania: change is occurring but (much) more is needed

- **What needs to happen to reform standard practices with P-values and hypothesis testing**
- **Connections to broader issues in statistical practice**

39

Tackling dichotomania

Greater effort needed to:

- **Remove "statistical significance" from the lexicon (and beware coded versions, e.g. "no difference was found")**
- **De-emphasise hypothesis testing as the primary statistical activity**
  - Emphasise interval estimation ("compatibility interval"?)
  - … and perhaps more Bayesian inference
- **Emphasise that inference in the face of random variation is uncertain, not dichotomous**
  - Remember all the other "hypotheses" (beyond the null) that may be of interest (leading back to estimation of effect)
  - View/present results as incremental information about effect sizes, not "findings" of effects present or absent

40

ASIDE: It's not the P-value per se that's the problem, it's the dichotomisation as "significant/ non-signif."

- If we no longer use *P*-values to declare "significance", what use are they?
  - Need to pay closer attention to what they actually mean…
- Valid interpretation is not easy: it seems very natural to try and convert the "probability" involved into a statement of uncertainty about the test/null hypothesis ("inversion fallacy")
- Rafi & Greenland (2020) revive a proposal to assist with this by using an equivalence between *P*-values and so-called *S*-values…

41

41

Rafi & Greenland (2020)

- *P*-value as measure of *compatibility* of data with test hypothesis (and assumptions)
- Also need to consider compatibility with other hypotheses
- NO SHARP THRESHOLDS ("significance")
- Propose *S*-value as alternative, safer measure of compatibility or "surprise"
  - Analogy with coin toss: how much "surprise" information is there in seeing *n* heads in a row?

42

42

"*S*-values" (Rafi & Greenland, 2020)

***S* for surprise, not significance!**

1 Head	**50-50**
2 Heads	**$(1/2)^2 = 0.25$**
3 Heads	**$(1/2)^3 = 0.125$**
4 Heads	**$(1/2)^4 = 0.063$**
5 Heads	**$(1/2)^5 = 0.031$**
…	
10 Heads	**$(1/2)^{10} = 0.001$**

P = 0.05 is as surprising (under the test hypothesis) as getting between 4 and 5 Heads when tossing a fair coin

43

43

Connections with other misuses and misinterpretations of statistical methods

- **Ritual (or at least habitual) application of techniques**
- **Regression rules the world**
  - *I see a binary outcome, therefore I fit a logistic regression…*
- **Methods producing answers in search of questions**
- **Overarching solution to all of these problems is to go back to the question(s)…**
  - If we recommend "estimation not testing", are we clear about what we are estimating, beyond a statistically defined "effect"?

44

44

The scourge of multivariable regression

- **Review shows many applications of regression methods are poorly framed:**
  - "adjustment" for covariates without clear causal intent
  - "Type 2 fallacy": interpretation of multiple coefficients in single model
  - "exploratory" use of regression to identify "risk factors" (no clear research question)
  - "prognostic factor" studies claiming to identify "independent effects"

FEATURED ARTICLE OPEN ACCESS

On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching

John B. Carlin[1,2,3] | Margarita Moreno-Betancur[1,2]

[1]Clinical Epidemiology & Biostatistics Unit (CEBU), Murdoch Children's Research Institute, Melbourne, Victoria, Australia | [2]CEBU, Department of Paediatrics, The University of Melbourne, Melbourne, Victoria, Australia | [3]Centre for Epidemiology & Biostatistics, Melbourne School of Population & Global Health, University of Melbourne, Melbourne, Victoria, Australia

Correspondence: John B. Carlin (jbcarlin@unimelb.edu.au)

Received: 12 September 2023 | **Revised:** 3 September 2024 | **Accepted:** 26 September 2024

Keywords: biostatistics training | epidemiological methods | misuse of statistics | regression models | three types of question

Statistics in Medicine (2025) with accompanying commentaries and authors' rejoinder

A grand unifying solution!

First, clarity of purpose: research question either descriptive, predictive or causal

- Trials are always asking a causal question

From type of question flows desired endpoint of analysis: what is the estimand? (the quantity in the infinite target population that represents the answer to the question)

- Fits with recent emphasis on estimands in trials: specify the target quantity before the analysis plan!
- N.B. Is it really of interest to ask the question "is the causal effect of treatment X positive ('non-null')?"?

Once estimand is clear, assess whether data can (in principle) provide a valid estimate (measurement issues, adherence, missing data, etc), then specify method for *estimation*

A grand unifying solution!

- **Proposed "roadmap" for answering questions has no place for testing hypotheses**
- **Instead, sharp focus on estimating a quantity of meaningful interest**
- **This approach helps to facilitate the "estimation not testing" mantra, by clarifying the estimand(s)**
  - No longer focusing on parameters in statistical models as the "effects of interest"
  - Models will be useful but need to be deliberately constructed with full articulation of assumptions that link parameters to estimand(s) of interest

See Carlin & Moreno-Betancur, *Stat Med* (2025) for more

A final word on the null hypothesis

"Statisticians classically asked the wrong question—and were willing to answer with a lie. They asked 'Are the effects of A and B different?' and they were willing to answer 'no.'

All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'are the effects different?' is foolish."

John W. Tukey, "The Philosophy of Multiple Comparisons", *Statistical Science* (1991)

http://xkcd.com/892/

It's time to take this great statistician's thoughts seriously!