

p -values: beyond polemics to practical proposals

James Carpenter

London School of Hygiene & Tropical Medicine

James.Carpenter@lshtm.ac.uk

Thu 1st Oct 2026, ISCB Freiburg

Acknowledgements



Michal Abrahamowicz (McGill, Canada)

Els Goetghebeur (Ghent, Belgium)

Willi Sauerbrei (Freiburg, Germany)

Carsten Schmidt (Greifswald, Germany)

Victor Kipnis (NIH, USA)

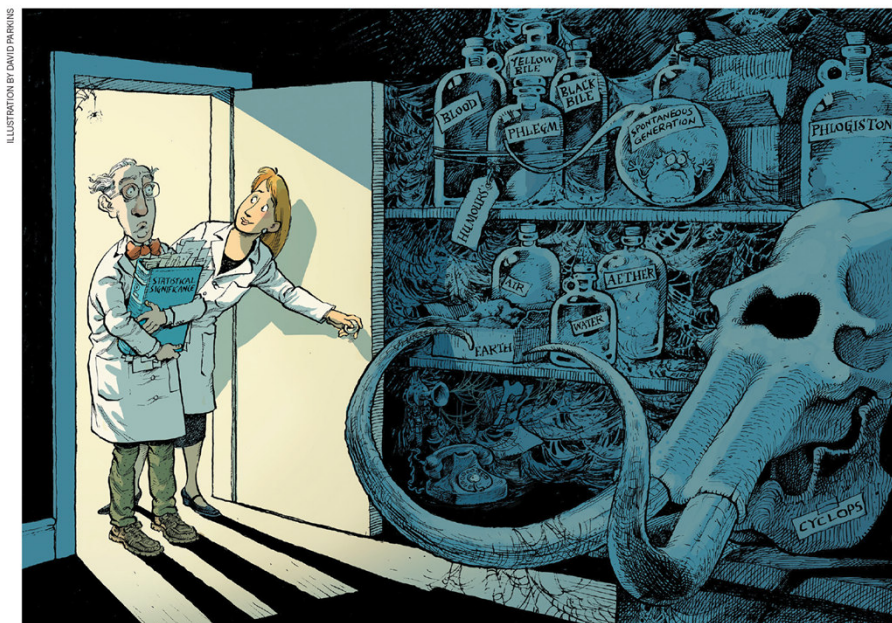
on behalf of the STRATOS p-values working group

Overview



- Criticism of use and reporting of p-values
- Reproducibility and p-values
- Use of p-values in three broad modelling aims
 - Descriptive
 - Predictive
 - Explanatory / Causal
- Examples: descriptive analysis; explanatory / causal analysis
- Discussion

Critique of p-values



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.

- Large literature on misuse of p-values
 - 2019 paper by AMR [1] arguing for retiring ‘statistical significance’
 - ‘confidence interval’ should be renamed to ‘compatibility interval’
 - Emphasize the need to focus on estimates and the uncertainty in them, including p-values
- Principal issue is lazy use and (deliberately?) misleading interpretation & reporting
- Well written, but other guidance needed

Some consequences of AGM



- Over-interpretation of point estimates may be an issue:
 - E.g. Pannikar *et al*, (2019) [2], Methods: “To avoid placing too much emphasis on statistical significance, we emphasize the strength of associations... as well [cite AGM].”
 - Then, Results section explicitly interprets OR = 5.456, with 95% CI: 0.3–90.6 as “5 times” risk increase
 - Re-constructed p-value is about 0.24
- Authors’ behaviour remains unchanged:
 - E.g. Marmor *et al* (2020) [3] cite AGM, but then report “significant associations” or “effects” in the conventional way:
‘... AI/AN women were found to be significantly more likely to have a high-risk (OR=1.28; 95% CI: 1.01-1.66)’

Can STRATOS help?



- STRATOS stands for Strengthening Analytical Thinking for Observational Studies
 - Hypothesis testing, inference and complete, transparent reporting are central to this
 - We aim to clarify practical issues
- Discussion highlighted concern that the AMR paper could be (mis)used by those who wish to jettison more formal inference
 - The current situation is a ‘curate’s egg’: but it could be worse!
- Therefore, we agreed to develop a manuscript which aimed to provide more practical guidance
 - We concluded the most appropriate way was to illustrate the principled use of p-values, testing and statistical significance in carefully chosen examples

Reproducibility and p-values



Reproducibility is, rightly, an increasing focus of efforts to improve research. However, many researchers are not aware that the p-value is not a reproducible statistic.

For a sample mean from a population with known variance 1, the table shows the probability of the p-value from a second experiment being < 0.05 , if the true mean is equal to that observed in the first experiment:

<i>p-value from 1st experiment:</i>	0.05	0.01	0.005	0.001
Pr($p < 0.05$) in repeat of experiment:	0.5	0.73	0.8	0.91

Use of p-values & significance testing in research

Large phase-III randomised trials are used to decide whether to adopt new treatments – in this setting, formal hypothesis testing is broadly agreed to be appropriate: reproducibility comes from the design which controls the type-1 error and power.

Observational research has much broader aims, main aims are classified as [e.g., 4]:

- Descriptive modelling
- Predictive modelling
- Explanatory / Causal modelling

p-values also play a broader role in observational research

Using p -values for reproducible research:

Follow what AMR propose relating to

- Presenting p -values with parameter estimates and confidence intervals.
- Not interpreting $p \geq \alpha$ as evidence of the ‘lack of effect’ but, instead, as ‘insufficient evidence to reject the null hypothesis’.
- Not interpreting parameter estimates, confidence intervals, and p -values isolated from previous / related findings.
- Not comparing effect sizes between groups, or across studies, using p -values.

Focus on reporting both methodology (i.e. methodological decisions) and results so that analyses are truly reproducible (Resnik & Shamoo, (2017))[12].

Descriptive modelling - Example



- Understand the relationship between a continuous outcome and ten continuous and binary variables in a large dataset.
- Key tasks: select variables and select functional forms
 - Use pre-defined choice of tuning parameters to control both the number of variables, and the complexity of the functional form (including interactions)
 - One of the popular approaches is Multivariable Fractional Polynomial (MFP) modelling. Here, p-values are the tuning parameters [8]

p-values as tuning parameters



- The MFP procedure uses two tuning-p-values:
 1. for variable selection: this controls how strong a variable's (potentially non-linear) effect needs to be for inclusion;
 2. To select functional form (functional selection procedure): are we interested to identify small deviations from linearity (large p-value) or would it be acceptable to oversee smaller deviations (small p-value) and specify a simpler model with a linear function.

For the multivariable analysis, MFP starts with the full model and combines backward elimination with the functional selection procedure for continuous variables.

Comments I:



- These p-values are not sacred ‘p-value thresholds’ but model-complexity tuning parameters.
- The selected model is not ‘the final truth’ but the primary selected MFP model.
- Sensitivity and stability analyses are important:
 - Sensitivity analysis: using different p-values to investigate model-complexity
 - Stability analyses: bootstrap procedures could be used
- It is important that all steps of this process are reported, ideally by following reporting guidelines, e.g. STROBE [6].

Comments II:

- Ideally scientific knowledge is used to pre-specify a model.
- However, in most analyses, data dependent decisions are still needed, although parts of the model may be 'known'. For the 'unknown' part, tuning p-values play an important role for model building.
- It is often argued that p-values should be replaced by information criteria (eg AIC, BIC) but that does not solve the issue. For example, for a one degree of freedom term, 0.157 corresponds approximately to AIC. p-values allow suitable choices according to the main aim of an analysis.
- Emphasis needs to be on improved methodological reporting (e.g. [7]), rather than debates over the utility of p-values.

Explanatory / Causal analysis: pre-specify

With observed exposures and focus on target trial emulation [9, 12] and causal estimands, safeguarding the causal null is still critical, but now also hinges on (untestable) causal assumptions, e.g. involving instrumental variables or no-unmeasured confounding [11].

These assumptions are captured in the DAG. To defend the conditional independence of exposure and outcome shown in the DAG (whether the analysis use adjusted outcome models or propensity scores) simple p-values are less helpful. This is because confounders draw on their *joint* adjusted association with exposure & outcome.

However, p-values provide an informal distance metric when checking covariate balance and comparing their means among the exposed and unexposed groups.

Discussion



- P-values are not reproducible: specifying type-1 error and power control this.
- Point estimates, confidence intervals and p-values are all indispensable inferential tools. Usually, all three should be reported.
- P-values can also have a distinct use as tuning parameters in model building.
- Reproducibility is a key characteristic of good research. To support this, we need
 - Increased focus on registration of studies
 - Publication of study protocols, including statistical analysis plans specifying how p-values will be used and interpreted
 - Improved methodological reporting (as well as reporting of results)
- We will give illustrative examples in the paper as these are a key tool to change practice.
- Journals have a key role to play –
papers with poorly reported methods should be rejected without review.

Selected references



1. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature* 2019;567:305-307
2. Panikkar B, Lemmond B, Allen L, Dipirro C, Kasper S. Making the invisible visible: Results of a community-led health survey following PFAS contamination of drinking water in Merrimack, New Hampshire. *Environmental Health: A Global Access Science Source* 2019;18(1):79.
3. Marmor S, Longacre CF, Altman AM, Hui JYC, Jensen EH, Tuttle TM. Genomic expression assay testing among American Indian and Alaska Native women with breast cancer. *Cancer* 2020;126(24):5222-5229.
4. Carlin JB, Moreno-Betancur M. On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching. *Stat Med.* 2025 Jun;44(13-14):e10244. doi: 10.1002/sim.10244. PMID: 40553044; PMCID: PMC12186762.
5. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative.
6. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)statement: guidelines for reporting observational studies. *Lancet.* 2007 Oct 20;370(9596):1453-7. PMID: 18064739
7. Sauerbrei, W., Ambrogi, F., De Bin, R., Boulesteix, A. L., Goetghebeur, E., & Huebner, M. (2025). Commentary: Regression Models—Efforts Are Required to Improve Statistical Practice and Teaching. *Statistics in Medicine*, 44(13-14), e10341.
8. Sauerbrei, W., Royston, P., Kipruto, E. (2025). Multivariable Fractional Polynomial Models and Extensions. In: Lovric, M. (eds) *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-69359-9_397
9. Ahlem Fourati, Marion Vergonjeanne, Thibaut Gellé, Julien Magne, A practical guide to target trial emulation: connecting randomized trials and real-world data in cardiovascular research, *European Journal of Preventive Cardiology*, 2026;, zwag267, <https://doi.org/10.1093/eurjpc/zwag267>
10. Swanson S, Labrecque J, and Hernán M. Causal null hypotheses of sustained treatment strategies: What can be tested with an instrumental variable? *European Journal Of Epidemiology* 2018; 33(8): 723-728
11. Hansford HJ, Cashin AG, Jones MD, et al. Reporting of observational studies explicitly aiming to emulate randomized trials: a systematic review. *JAMA Netw Open.* 2023;6:e2336023. [PMID: [37755828](https://pubmed.ncbi.nlm.nih.gov/37755828/)] doi: [10.1001/jamanetworkopen.2023.36023](https://doi.org/10.1001/jamanetworkopen.2023.36023)
12. Resnik, D. B., & Shamoo, A. E. (2017). Reproducibility and research integrity. *Accountability in research*, 24(2), 116-123.