Medicine Is Being Treated with Snake-Oil Statistics

Zad Rafi[a], Andrew Gelman[b], Aleksi Reito[c], Sander Greenland[d]

[a]*Department of Population Health, NYU Langone Medical Center, New York, NY, USA,*

[b]*Department of Statistics and Department of Political Science, Columbia University, New York, NY, USA,*

[c]*Department of Surgery, Central Finland Hospital, Jyväskylä, Keski-Suomi, Finland,*

[d]*Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA,*

Suggested Running Head: Snake Oil Statistics

Statistics has helped medicine move away from an eminence-based framework, where subject-matter experts decided what worked and what didn't, towards an evidence-based one. Among the techniques deployed, null-hypothesis statistical testing (NHST) is the most common, where the word "null" is invariably taken by users to mean that the only hypotheses tested are those of "no association" or "no effect" (sometimes labeled "nil hypotheses"). Despite objections to them extending throughout the past century [1], these tests are supposed to serve as a safeguard against researchers fooling themselves and others, as well as serving as a central component of experimental design and analysis.

Unfortunately, these tests are easily subverted into playing the reverse role of providing a badge of approval, allowing researchers to make stronger claims than warranted by valid statistical analyses. Confidence intervals have been extensively promoted to address this problem, but they too have been subverted by being treated as if they are only testing null hypotheses. The consequences for evidence-based medicine have been dire.

There are three commonly stated principles (shown in Figure 1) of evidence-based medicine [2]:

1. reliance on statistically significant results (and thus NHST) from randomized controlled trials,
2. balancing of costs, benefits, and uncertainties in decision making, and
3. combining clinical expertise with external evidence to tailor treatments for individuals.

*Corresponding author

Figure 1

Unfortunately, the use of NHST can get in the way of the movement toward an evidence-based framework. This may sound paradoxical, given that one of the foundations of evidence-based medicine is hypothesis testing based on randomized controlled trials, deemed by many to be the most reliable forms of evidence [2]. One problem is that reliance on statistical significance (principle 1) may conflict with the other principles: In conflict with balancing costs and benefits under uncertainty (principle 2), statistical significance or non-significance is typically used to replace uncertainty with certainty [3, 4] — indeed, researchers are encouraged to do this and it is even forced by some journals [5].

Reliance on statistical significance can interact badly with expertise and background evidence (principle 3), often leading to incoherent attempts at resolution. For example, researchers may do separate analyses for men and women, often with no biological basis for expecting other than small differences in effects (if any). In doing so they often find that one group (usually the larger one, which is usually men) show a "significant" effect while the other does not. They then misreport this difference in "significance" as if it represented a significant difference in effect between groups when in fact it only reflects a difference in the group sizes, plus random differences in group-specific P-values. Such misinterpretations fool researchers into thinking that the data support targeting treatments at the group showing "significance" and mislead clinicians into tailoring treatment plans for individuals based on nothing more than random variation.

The unfortunate reality is that estimating effects for individuals or population subsets requires far more sophistication than basic testing procedures [6]. It can take several times more patients to estimate variation in effects ("interactions") than average effects [7]. Given that few studies are large enough to estimate main effects of interest, it will typically be impossible to obtain reliable estimates of effect variation even if the study is otherwise flawless. That problem should be dealt with under principle (2) by recognizing that subgroups will have very imprecise estimates, leading to enormous uncertainties about who if anyone should be targeted for or excluded from treatment. Unfortunately, the prevalent misunderstandings of classical statistics have trained most researchers, editors, and reviewers to demand statistical significance and certainty as a prerequisite for publication [8] and decision making. Applying those demands within subgroups all but guarantees that distorted impressions of patient-specific effects will follow.

Through neglect of basic education, the statistics profession is partly responsible for these issues, but there is nothing new about medicine's desire for certainty. That desire is part of human nature and thus existed long before the adoption of statistical methods. The resulting demands for certainty in reported results have provided ample opportunities for overconfident researchers to rise to prominence. Physicians have long capitalized on several opportunities with their sophisticated knowledge of physiology and anatomy, and used their medical authority to argue what treatments worked, shape public policy, and design clinical guidelines, maintaining a form of medicine characterized as "eminence based."

The landscape changed when advances in quantitative methods eventually reached medicine [9], leading to a newfound demand for scientific rigor and making many individuals wary of the claims of subject-matter experts [10]. These demands may be some of the biggest contributors to medicine's adoption of statistical methods and its movement towards an evidence-based framework. Unfortunately, in its pursuit for objectivity, medicine was sold snake oil statistics and as a result, its desire for objectivity and rigor backfired.

Figure 2: Snake oil statistics. Methodological interventions that have been oversold by early adopters as cure alls despite no supporting evidence of utility in a particular application.

To see the reach of these problems, one simply needs to look wherever a new drug has been considered effective only if it has been shown statistically significantly better than a control in one or more randomized clinical trials (a standard that applies to FDA new-drug approvals, though not to medical devices, which can be cleared through other pathways). The largest enforcers of these methods have been regulatory agencies that wish to minimize treatments that do not work (false positives) and minimize adverse events from treatments. One should ask: what specifically convinced these agencies that statistical significance and in particular NHST was the best analysis criterion to achieve these goals?

In a nonexistent ideal world, the regulatory agencies looked at several statistical methods, tested each of them in many settings to see how well they identified true benefits and harms while avoiding false conclusions, and from that determined NHST performed best out of all options – with periodic updates as new methods appeared. Of course, history shows something else entirely, even indicating that adoption of NHST was mainly a result of political desperation [11, 12]. To understand the harsh reality, we must go back to the mid-20th century, when pharmaceutical companies submitted new drug applications to the FDA that often lacked any protocols and statistical analysis plans, making the entire drug approval process chaotic. By the end of the 1960s, the FDA had become desperate to standardize the drug-approval process and make it scientifically rigorous, especially with impending pressures from the Drug Amendments of 1962, which demanded rigorous evidence for drug approval.

At the same time, applied researchers were looking for rigorous ways to summarize experiments with numerical quantities. Their search struck gold with the works of the prominent statisticians, Ronald A. Fisher [13, 14, 15], Jerzy Neyman [16], and Egon Pearson [17], who created and popularized powerful statistical tools for researchers that lacked statistical training. Eventually, researchers across many disciplines adopted NHST and the now-infamous 0.05 cutoff. Soon, the FDA followed and incorporated these methods into its regulatory process, without any formal debate about the methods' utility or evidence [12].

Figure 3: Austin Bradford Hill. The statistician and epidemiologist who conducted the first randomized controlled trial in medicine.

It thus appears that embracement of the NHST paradigm arose from expediently

following the ascendant trend in research rather than a critical evaluation of various options emerging in the same period. Since its adoption into the regulatory process, this framework has been rigorously enforced by the FDA as the gold standard in the approval process.

We find it ironic that the gatekeepers of evidence-based medicine — regulatory agencies, journal editors and reviewers, and medical researchers – continue to insist on enforcement of a statistical framework that was never critically examined for its utility in medicine [12]. This failure makes it understandable why it can be credibly argued that conventional statistical methods have done more harm than good in medical science. These methods, including null-hypothesis tests and so-called "confidence" intervals are a set of decision-making tools designed for tightly controlled randomized experiments in which the treatment effects (if any) can be distinguished from all other causal effects, and can be distinguished from random error (noise) by increasing the study size in an affordable manner. Examples of such scenarios arise in agricultural research where the experimental units are plants or plots, and industrial quality control, where the experimental unit is a part or product. The tools were originally developed for these environments, in which (compared to clinical studies) experimenters have almost godlike control over the selection of units and their subsequent experiences, aided by having a rather short follow-up period and low cost per experimental unit. And then, the decisions to be made are relatively simple and easily monitored, e.g., change a fertilizer formulation or manufacturing tolerance [17]. Experimental psychology is similar in terms of control, cost, and even lower cost-benefit consequences.

In clinical environments, however, the cost per experimental unit (the patient) is far higher, creating severe limits on study size and thus noise reduction, and the costs and benefits of decisions can be enormous. At the same time, there is far less control of extraneous selection and confounding effects: Physicians and their patients can and do selectively refuse to participate or cease to adhere to assigned treatment protocols, and may drop out for unknown reasons. Meanwhile, direct physical control of the patient environment is extremely limited or nonexistent, especially when follow-up extends beyond hospital stay. And then, amplifying these limits, the required decisions are often complex and of highly uncertain consequence, yet may be pivotal to the experimental results and clinical decisions, e.g., when to withdraw treatment from patients apparently experiencing side effects or when to switch treatments for nonresponsive patients.

These vast differences haven't stopped medical researchers from using conventional methods as if they were operating in a tightly controlled environment, treating ambiguous trial results as if they supported decisive conclusions despite obvious uncertainties and potentially devastating consequences. The usual depictions of this problem involves researchers trying to "game" statistical methods so that they show "significant" effects [18, 19]; while such "significance questing" is a real problem, warnings about it usually ignore or dismiss opposite behavior in which showing "nonsignificance" will facilitate publication in prestigious medical journals, especially where so-called "replication failure" has become a

hot topic [3]. Adding to these distortions is the publication bias that results when researchers or editors deem results unworthy of submission or acceptance because they are just not interesting because they fail to report a "discovery" or only replicate what is "known." This desire for novelty or publicity, along with demands for certainty, has distorted the medical literature with dubious, inflated effects [20] and misleading claims of replication failure based on fundamentally ambiguous results [21, 22].

0.1. Responses to the problems

In response to the ongoing abuse of statistical testing, the American Statistical Association released a statement in 2016 cautioning against the fixating on P-values and statistical significance [23]. Three years later, the organization published an issue titled "A World Beyond $P < 0.05$" with 43 commentaries from statistical experts on how to improve statistical inference, with or without P-values [24]. The issue was accompanied by a highly discussed commentary in Nature titled "Scientists Rise up Against Statistical Significance" that discouraged mindless automation and dichotomization of statistical results [21], and was supported by signatures of some 800 applied researchers.

While these calls are impressive and a growing number of journals — including the New England Journal of Medicine and, in modified form, JAMA — have updated their statistical-reporting policies, most regulatory agencies and the bulk of the medical literature continue to fixate on statistical significance [5, 25]. Correspondingly, we can expect medical researchers to keep cutting corners to achieve or remove statistical significance [18, 19] or misinterpret ambiguous results as if definitive, a practice that is often labeled as "spin" [26].

We can interpret the persistence of null hypothesis significance testing (NHST) in two ways. One story is that the value of NHST is recognized by real-world decision makers, despite the carping of ivory-tower critics such as the authors of the present article. The other story is that the counterproductive nature of NHST has been denied by the medical establishment [5, 25], and that better alternatives are available [12]. There are legitimate practical concerns behind both these perspectives. On one hand, active researchers and regulators have legitimate concerns about working on or approving treatments that do not work. On the other hand, many well-publicized examples have made it clear that NHST can easily lead to overconfidence and erroneous inferences, as seen in discussions that treat statistically significance as demonstrating presence of effects and nonsignificance as demonstrating absence of effects.

0.2. Methodologists propose new methods to address NHST in their field

Fear of false positives has dominated most discussions of scientific rigor [27, 28, 18], often relegating false negatives to more technical discussions of power. Although false positives can be costly, so can missing clinically meaningful effects [21]. Take postmarketing surveillance for pharmaceuticals; in such situations, serious adverse events from drugs are often underreported and have in some cases

been actively suppressed [29, 30], as seen in Cochrane reviews of unpublished trial data, leading to a scarcity of data and resulting in studies that will never be able to show a "significant" effect because of the lack of resources and time. If a study comparing adverse events from those taking an approved drug and some control group is unable to show statistical significance, even when the estimated effect is important and plausible, the results will typically be confused and used for evidence of absence [31, 21], and prescribing is unlikely to be curtailed, leading to continued harm.

The confusion of statistical nonsignificance with evidence of absence will remain a problem in fields where effects are often small and yet studies powered to detect them are infeasible. Most studies in surgical science rarely have more than a dozen participants. Again, this makes it incredibly difficult to achieve statistical significance, even when clinically important effects are likely. In such fields, adopting the NHST framework for analysis all but guarantees failure to detect those effects. As a result, researchers have looked to other methods — including Bayes factors, second-generation P-values, and post-hoc power calculations — to circumvent the bad hand that was dealt to them [32, 33, 34]. Unfortunately, such transitions are not always helpful: while reasonably defensible alternatives such as interval testing and posterior intervals do exist, several of the most-publicized proposals introduce their own errors, and in fact, some may actually result in more errors than before.

For example, a group of surgeons recently published several statistical recommendations on what to do if a surgical study result was nonsignificant [35, 36, 37], providing hope to researchers who have had their inquiries halted by that result. But the recommendations are not only statistically invalid and clinically misleading, and thus have been discouraged by statisticians for nearly two decades [38, 39]. In a similar tale, two sports scientists published a statistical method [40] which was supposed to improve the classification of individual treatment responses by reducing the influence of random error in small studies. Unfortunately, reducing the influence of random error in a valid manner requires improvement of study design, including increase in study size, so unsurprisingly the proposed method has unacceptably poor statistical properties [41].

The efforts behind new methods and recommendations are a response to an arbitrary dichotomy that has been imposed upon researchers by the medical and scientific establishment: decide whether a result is "positive" or "negative", with no allowance for what is usually the most reasonable interpretation – ambiguous or indecisive. Unfortunately, conventional statistics has aggravated this problem by offering methods like NHST which produce only dichotomous answers. These methods which swept research sciences in the mid-20th century and are now firmly rooted in tradition, as if that tradition is the best we can do. But it isn't.

The problem with upending this tradition however is that there is no consensus about what to replace it with, a problem that only grows as alternatives continue to proliferate. A idealistic (and we think naïve) view would simply allow authors to choose alternatives of their liking. The problem with this anarchic approach

is that many of the alternatives are themselves misleading and even defective or at best inferior to other methods in demonstrable ways. Yet seeing these problems requires not only sufficient technical expertise to evaluate methods, but also a willingness to see them – a willingness that should not be assumed for originators and adopters of the method. Unfortunately, mere publication of a method in a peer-reviewed journal is a faulty indicator of method validity. That is especially so if publication is not in a statistics journal, for in that case it means only that peer-review was by referees who may have had little of the special technical expertise needed for thorough evaluation.

0.3. Solutions

There are no simple solutions or universal guidelines that medical researchers can always use to improve scientific rigor within their area of work. We nonetheless offer some recommendations and resources that we believe may be useful to those who recognize problems in their field and wish for some sort of guidance:

1. Collaborate with well-qualified statisticians to design and conduct studies that are rigorous, efficient, and cost effective [42, 43]. Look to these statistical collaborators for guidance about honestly reporting uncertainty, not certainty [19].
2. If possible, pool resources to run larger studies that are likely to be more precise and informative than individual studies, which may simply waste resources and offer little yield [44]. However, more data is not synonymous with more information: Increasing study size may be detrimental if it entails a reduction in data quality [45].
3. Aim to be more descriptive and less inferential [22]. Accept uncertainty and the anxiety that comes with it [4], along with the idea that no one study warrants conclusions about the true nature of a phenomenon, a delusion that is not even true in particle physics! And if a phenomenon is subtle, even several studies may be insufficient for valid conclusions beyond "more research is needed."
4. When making real world decisions, use all information available to you, balancing costs, benefits, and uncertainties [46], rather than basing decisions on whether a single numerical value is above or below an arbitrary cutoff.
5. Be mindful of cognitive biases that may distort your conclusions throughout the study and the many cognitive biases that will afflict you, your colleagues, and your collaborators [47, 3, 48].
6. Use statistical methods that have been reviewed and validated by the statistical community beyond their developers and promoters. That validation is provided not only by publication in statistical journals, but also by applications that can be judged as having reached sound, well-cautioned conclusions in context. Especially, beware of any method that (like NHST) claims to offer firm conclusions based on purely numeric comparisons.

References

[1] E. G. Boring, Mathematical vs. scientific significance, Psychological Bulletin 16 (10) (1919) 335–338, https://doi.org/10.1037/h0074554. doi:10.1037/h0074554.

[2] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, W. S. Richardson, Evidence based medicine: What it is and what it isn't, BMJ (Clinical research ed.) 312 (7023) (1996) 71–72. doi:10.1136/bmj.312.7023.71.

[3] S. Greenland, Invited commentary: The need for cognitive science in methodology, American Journal of Epidemiology 186 (6) (2017) 639–645, https://doi.org/10.1093/aje/kwx259. doi:10.1093/aje/kwx259.

[4] B. B. McShane, D. Gal, A. Gelman, C. Robert, J. L. Tackett, Abandon Statistical Significance, The American Statistician 73 (sup1) (2019) 235–245, https://doi.org/10.1080/00031305.2018.1527253. doi:10.1080/00031305.2018.1527253.

[5] H. Bauchner, R. M. Golub, P. B. Fontanarosa, Reporting and Interpretation of Randomized Clinical Trials, Journal of the American Medical Association 322 (8) (2019) 732–735, https://doi.org/10.1001/jama.2019.12056. doi:10.1001/jama.2019.12056.

[6] S. Senn, Statistical pitfalls of personalized medicine, Nature 563 (7733) (2018) 619–621, http://www.nature.com/articles/d41586-018-07535-2. doi:10.1038/d41586-018-07535-2.

[7] S. Greenland, Tests for interaction in epidemiologic studies: A review and a study of power, Statistics in Medicine 2 (2) (1983) 243–251, http://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780020219. doi:10.1002/sim.4780020219.

[8] R. Rosenthal, The file drawer problem and tolerance for null results, Psychol. Bull. 86 (3) (1979) 638, http://psycnet.apa.org/record/1979-27602-001. doi:10.1037/0033-2909.86.3.638.

[9] M. R. Council, Streptomycin Treatment of Pulmonary Tuberculosis, British Medical Journal 2 (4582) (1948) 769–782, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2091872/.

[10] S. N. Goodman, Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics, The American Statistician 73 (sup1) (2019) 26–30, https://doi.org/10.1080/00031305.2018.1558111. doi:10.1080/00031305.2018.1558111.

[11] L. Kennedy-Shaffer, When the Alpha is the Omega: P-Values, "Substantial Evidence," and the 0.05 Standard at FDA, Food Drug Law J. 72 (4) (2017) 595–635, https://www.ncbi.nlm.nih.gov/pubmed/30294197.

[12] S. J. Ruberg, F. E. H. Jr, M. Gamalo-Siebers, L. LaVange, J. J. Lee, K. Price, C. Peck, Inference and Decision Making for 21st-Century Drug Development and Approval, The American Statistician 73 (sup1) (2019) 319–327, https://doi.org/10.1080/00031305.2019.1566091. doi:10.1080/00031305.2019.1566091.

[13] R. Fisher, Statistical Methods and Scientific Induction, Journal of the Royal Statistical Society. Series B (Methodological) 17 (1) (1955) 69–78, http://www.jstor.org/stable/2983785. doi:10.1111/j.2517-6161.1955.tb00180.x.

[14] R. A. Fisher, Statistical Methods for Research Workers, Edinburgh, Oliver and Boyd, 1925, https://books.google.com/books?id=GmNAAAAAIAAJ&q.

[15] R. A. Fisher, The Design of Experiments, The Design of Experiments, Oliver & Boyd, Oxford, England, 1935.

[16] J. Neyman, E. S. Pearson, On the Problem of the Most Efficient Tests of Statistical Hypotheses, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231 (1933) 289–337, http://www.jstor.org/stable/91247. doi:10.1098/rsta.1933.0009.

[17] E. S. Pearson, A Survey of the Uses of Statistical Method in the Control and Standardization of the Quality of Manufactured Products, Journal of the Royal Statistical Society 96 (1) (1933) 21–75, http://www.jstor.org/stable/2341869. doi:10.2307/2341869.

[18] J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant, Psychological Science 22 (11) (2011) 1359–1366, https://doi.org/10.1177/0956797611417632. doi:10.1177/0956797611417632.

[19] M. Q. Wang, A. F. Yan, R. V. Katz, Researcher requests for inappropriate analysis and reporting: A U.S. survey of consulting biostatisticians, Annals of Internal Medicine 169 (8) (2018) 554, https://doi.org/10.7326/M18-1230. doi:10.7326/M18-1230.

[20] K. Dickersin, J. A. Berlin, Meta-analysis: State-of-the-Science, Epidemiologic Reviews 14 (1) (1992) 154–176, https://doi.org/10.1093/oxfordjournals.epirev.a036084. doi:10.1093/oxfordjournals.epirev.a036084.

[21] V. Amrhein, S. Greenland, B. McShane, Scientists rise up against statistical significance, Nature 567 (7748) (2019) 305, https://doi.org/10.1038/d41586-019-00857-9. doi:10.1038/d41586-019-00857-9.

[22] V. Amrhein, D. Trafimow, S. Greenland, Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication, The

American Statistician 73 (sup1) (2019) 262–270, https://doi.org/10.1080/00031305.2018.1543137. doi:10.1080/00031305.2018.1543137.

[23] R. L. Wasserstein, N. A. Lazar, The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician 70 (2) (2016) 129–133, http://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108. doi:10.1080/00031305.2016.1154108.

[24] R. L. Wasserstein, A. L. Schirm, N. A. Lazar, Moving to a world beyond "p < 0.05", The American Statistician 73 (sup1) (2019) 1–19, https://doi.org/10.1080/00031305.2019.1583913. doi:10.1080/00031305.2019.1583913.

[25] D. Harrington, R. B. D'Agostino, C. Gatsonis, J. W. Hogan, D. J. Hunter, S.-L. T. Normand, J. M. Drazen, M. B. Hamel, New guidelines for statistical reporting in the journal, New England Journal of Medicine 381 (3) (2019) 285–286. doi:10.1056/NEJMe1906559.

[26] M. S. Khan, N. Lateef, T. J. Siddiqi, K. A. Rehman, S. Alnaimat, S. U. Khan, H. Riaz, M. H. Murad, J. Mandrola, R. Doukky, R. A. Krasuski, Level and Prevalence of Spin in Published Cardiovascular Randomized Clinical Trial Reports With Statistically Nonsignificant Primary Outcomes: A Systematic Review, JAMA Network Open 2 (5) (2019) e192622–e192622, http://jamanetwork.com/journals/jamanetworkopen/fullarticle/2732330. doi:10.1001/jamanetworkopen.2019.2622.

[27] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. H. Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, V. E. Johnson, Redefine statistical significance, Nature Human Behaviour 2 (1) (2017) 6–10, https://doi.org/10.1038/s41562-017-0189-z. doi:10.1038/s41562-017-0189-z.

[28] J. P. A. Ioannidis, The Proposal to Lower P Value Thresholds to .005, JAMA 319 (14) (2018) 1429–1430, http://dx.doi.org/10.1001/jama.2018.1536. doi:10.1001/jama.2018.1536.

[29] J. P. Vandenbroucke, B. M. Psaty, Benefits and risks of drug treatments: How to combine the best evidence on benefits with the best data about adverse effects, JAMA 300 (20) (2008) 2417–2419. doi:10.1001/jama.

2008.723.
URL https://doi.org/10.1001/jama.2008.723

[30] P. Doshi, T. Jefferson, C. Del Mar, The imperative to share clinical study reports: Recommendations from the Tamiflu experience, PLoS Medicine 9 (4) (2012) e1001201. doi:10.1371/journal.pmed.1001201.
URL https://doi.org/10.1371/journal.pmed.1001201

[31] D. G. Altman, J. M. Bland, Absence of evidence is not evidence of absence, BMJ (Clinical research ed.) 311 (7003) (1995) 485, https://doi.org/10.1136/bmj.311.7003.485. doi:10.1136/bmj.311.7003.485.

[32] G. Gigerenzer, J. N. Marewski, Surrogate Science: The Idol of a Universal Method for Scientific Inference, Journal of Management 41 (2) (2015) 421–440, https://doi.org/10.1177/0149206314547522. doi:10.1177/0149206314547522.

[33] D. van Ravenzwaaij, R. Monden, J. N. Tendeiro, J. P. A. Ioannidis, Bayes factors for superiority, non-inferiority, and equivalence designs, BMC Medical Research Methodology 19 (1) (2019) 71, https://doi.org/10.1186/s12874-019-0699-7. doi:10.1186/s12874-019-0699-7.

[34] U. Simonsohn, L. D. Nelson, J. P. Simmons, P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results, Perspect. Psychol. Sci. 9 (6) (2014) 666–681, http://dx.doi.org/10.1177/1745691614553988. doi:10.1177/1745691614553988.

[35] Y. J. Bababekov, S. M. Stapleton, J. L. Mueller, Z. V. Fong, D. C. Chang, A Proposal to Mitigate the Consequences of Type 2 Error in Surgical Science, Annals of Surgery 267 (4) (2018) 621, http://journals.lww.com/annalsofsurgery/Fulltext/2018/04000/A_Proposal_to_Mitigate_the_Consequences_of_Type_2.6.aspx. doi:10.1097/SLA.0000000000002547.

[36] Y. J. Bababekov, Y.-C. Hung, Y.-T. Hsu, B. V. Udelsman, J. L. Mueller, H.-Y. Lin, S. M. Stapleton, D. C. Chang, Is the Power Threshold of 0.8 Applicable to Surgical Science?—Empowering the Underpowered Study, Journal of Surgical Research 241 (2019) 235–239, http://www.sciencedirect.com/science/article/pii/S0022480419301854. doi:10.1016/j.jss.2019.03.062.

[37] Y. J. Bababekov, D. C. Chang, Post Hoc Power: A Surgeon's First Assistant in Interpreting "Negative" Studies, Annals of Surgeryhttps://doi.org/10.1097/SLA.0000000000002914 (Jan. 2019). doi:10.1097/SLA.0000000000002914.

[38] S. Greenland, Nonsignificance plus high power does not imply support for the null over the alternative, Annals of Epidemiology 22 (5) (2012) 364–368, https://doi.org/10.1016/j.annepidem.2012.02.007. doi:10.1016/j.annepidem.2012.02.007.

[39] J. M. Hoenig, D. M. Heisey, The Abuse of Power, The American Statistician 55 (1) (2001) 19–24, https://doi.org/10.1198/000313001300339897. doi:10.1198/000313001300339897.

[40] S. J. Dankel, J. P. Loenneke, A Method to Stop Analyzing Random Error and Start Analyzing Differential Responders to Exercise, Sports Medicinehttps://doi.org/10.1007/s40279-019-01147-0 (Jun. 2019). doi:10.1007/s40279-019-01147-0.

[41] M. Tenan, A. D. Vigotsky, A. R. Caldwell, On the Statistical Properties of the Dankel-Loenneke Method, SportRxivhttps://osf.io/8ndhg (Oct. 2019). doi:10.31236/osf.io/8ndhg.

[42] A. Gelman, J. Carlin, Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors, Perspectives on Psychological Science 9 (6) (2014) 641–651, https://doi.org/10.1177/1745691614551642. doi:10.1177/1745691614551642.

[43] K. J. Rothman, S. Greenland, Planning study size based on precision rather than power, Epidemiology 29 (5) (2018) 599–603, https://doi.org/10.1097/EDE.0000000000000876. doi:10.1097/EDE.0000000000000876.

[44] H. Moshontz, L. Campbell, C. R. Ebersole, H. IJzerman, H. L. Urry, P. S. Forscher, J. E. Grahe, R. J. McCarthy, E. D. Musser, J. Antfolk, C. M. Castille, T. R. Evans, S. Fiedler, J. K. Flake, D. A. Forero, S. M. J. Janssen, J. R. Keene, J. Protzko, B. Aczel, S. Á. Solas, D. Ansari, D. Awlia, E. Baskin, C. Batres, M. L. Borras-Guevara, C. Brick, P. Chandel, A. Chatard, W. J. Chopik, D. Clarance, N. A. Coles, K. S. Corker, B. J. W. Dixson, V. Dranseika, Y. Dunham, N. W. Fox, G. Gardiner, S. M. Garrison, T. Gill, A. C. Hahn, B. Jaeger, P. Kačmár, G. Kaminski, P. Kanske, Z. Kekecs, M. Kline, M. A. Koehn, P. Kujur, C. A. Levitan, J. K. Miller, C. Okan, J. Olsen, O. Oviedo-Trespalacios, A. A. Özdoğru, B. Pande, A. Parganiha, N. Parveen, G. Pfuhl, S. Pradhan, I. Ropovik, N. O. Rule, B. Saunders, V. Schei, K. Schmidt, M. M. Singh, M. Sirota, C. N. Steltenpohl, S. Stieger, D. Storage, G. B. Sullivan, A. Szabelska, C. K. Tamnes, M. A. Vadillo, J. V. Valentova, W. Vanpaemel, M. A. C. Varella, E. Vergauwe, M. Verschoor, M. Vianello, M. Voracek, G. P. Williams, J. P. Wilson, J. H. Zickfeld, J. D. Arnal, B. Aydin, S.-C. Chen, L. M. DeBruine, A. M. Fernandez, K. T. Horstmann, P. M. Isager, B. Jones, A. Kapucu, H. Lin, M. C. Mensink, G. Navarrete, M. A. Silan, C. R. Chartier, The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network:, Advances in Methods and Practices in Psychological Science (Oct. 2018). doi:10.1177/2515245918797607.

[45] D. R. Cox, C. A. Donnelly, Principles of Applied Statistics, Cambridge University Press, 2011, https://doi.org/10.1017/CBO9781139005036.

[46] G. Parmigiani, L. Inoue, Decision Theory: Principles and Approaches, John Wiley & Sons, 2009, https://doi.org/10.1002/9780470746684.

[47] G. Gigerenzer, Mindless statistics, The Journal of Socio-Economics 33 (5) (2004) 587–606, https://doi.org/10.1016/j.socec.2004.09.033. doi:10.1016/j.socec.2004.09.033.

[48] P. B. Stark, A. Saltelli, Cargo-cult statistics and scientific crisis, Significance 15 (4) (2018) 40–43, https://doi.org/10.1111/j.1740-9713.2018.01174.x. doi:10.1111/j.1740-9713.2018.01174.x.