The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics

Steven N. Goodman

To cite this article: Steven N. Goodman (2019) Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics, The American Statistician, 73:sup1, 26-30, DOI: 10.1080/00031305.2018.1558111

To link to this article: https://doi.org/10.1080/00031305.2018.1558111

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 26–30: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1558111

Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics

Steven N. Goodman

Departments of Medicine and Health Research and Policy, Stanford University, Stanford, CA

ABSTRACT

The current concerns about reproducibility have focused attention on proper use of statistics across the sciences. This gives statisticians an extraordinary opportunity to change what are widely regarded as statistical practices detrimental to the cause of good science. However, how that should be done is enormously complex, made more difficult by the balkanization of research methods and statistical traditions across scientific subdisciplines. Working within those sciences while also allying with science reform movements—operating simultaneously on the micro and macro levels—are the key to making lasting change in applied science.

ARTICLE HISTORY
Received April 2018
Revised November 2018

KEYWORDS
P-values; Reproducible research; Statistical inference, Scientific inference

My first statistical graduate advisor and mentor, Richard Royall at Johns Hopkins, warned that if I took on inference as a professional path, I would be lonely. He was not wrong, but the times have dramatically changed; I cannot recall a time in my professional career when so many statisticians and scientists have discussed issues in statistical inference with such energy and urgency. This is being produced by the largest metacontroversy that science may have yet faced—the "reproducibility crisis" —with one of its root causes being seen as misunderstanding and misuse of statistical methods. If we are interested in change, we cannot afford to waste a crisis.

To start to understand the trajectory going forward, it can be useful to see what its slope has been up to now. Consider this quote:

> …scientific inference from a set of data is not the formal exercise one finds taught in statistical classrooms. Today, the way one draws an inference from a real set of data is taught in many classrooms of statistics in exactly the same way as one would teach geometry or algebra. The student learns that statistical methods consist of a body of formulas and fixed sets of rules, which once memorized, can be used throughout one's lifetime in drawing inferences from data. We've learned one has only to determine whether to reject at the 5 per cent or 1 per cent level. Then the statistician can grandly draw obvious conclusions about data from any scientific field by proclaiming significance or non-significance. Such nonsense is taught usually by professors who have had minimal contact with the applications of statistical methods to scientific problems. As a result, the number of scientific papers which use statistical methods for window dressing is increasing. It appears that the P-value next to a contingency table is beginning to mean what the "Seal of Good Housekeeping" means …(Cutler et al. 1966).

Those are the words of Marvin Zelen in 1965, summarizing an NIH symposium on hypothesis testing in clinical trials, featuring talks from Sam Greenhouse, Jerry Cornfield, and Marvin Schneiderman. It is fair to say that the situation today is little different, and a major impetus of the ASA *P* value statement. If we are contemplating change, it is critical that we understand why things have improved so little over the last half century, and even since RA Fisher's warnings decades earlier. The answer is to be found more in the sociology and organization of science than in its technical or inferential foundations. I will focus mainly on the application of statistics within biomedical science.

Disciplinary Siloing of Methods

Almost all scientific methods are used, taught and communicated within disciplinary communities. It is within these communities that establish what will count for legitimate knowledge claims; cardiologist write for cardiologists, oncologists for oncologists, psychologists for psychologists, economists for economists. Each of these disciplines has somewhat different modes of experimentation and epistemic justification for recommendations. Most medical specialties look to randomized clinical trials on which to base clinical guidelines, whereas most surgical journals are filled with case series that rarely appear in the top journals of medical specialties.

Different disciplines have different historical memories about achievements and failures, like cultures that pass down archetypal stories that transmit values and customs to the next generation.

A clinical trial that shook the core of cardiology was the Cardiac Arrhythmia Suppression Trial (CAST), completed in 1991 (Echt et al. 1991). The therapies tested in the CAST trial were known to stop the arrhythmias thought responsible for

CONTACT Steven N. Goodman *steve.goodman@stanford.edu*
This essay was adapted from a keynote address given at the ASA Symposium on Statistical Inference, October, 2017.

sudden cardiac death, which was then killing more than one person every minute, or about 1200 per day. CAST revealed that these drugs, the most widely prescribed drugs in the United States, were almost quadrupling the sudden death rate, killing more Americans in the preceding decade than had died in most wars. This taught the cardiology community the danger of using surrogate endpoints, and the unreliability of mechanistic knowledge more effectively than could a thousand lectures, papers or causal diagrams. Yet few physicians in other disciplines know of it and have learned that lesson. These other areas have their own stories, yet even the most dramatic, like the Duke "omics" story (Micheel et al. 2012) are little known outside their own domains.

This siloing includes statistical methods. A methods paper in a cancer journal will usually be little read in another specialties; if it is in pediatrics, it will be little noticed in adult medicine, if it is in a surgical journal, not noted in medicine, and if in epidemiology, not read by most clinical investigators. In the mid-2000's many sports medicine researchers adopted a statistical method called "magnitude-based inference" (Batterham and Hopkins 2006). Subsequent papers showed the approach to be unfounded (Butson 2018; Sainani 2018), but it is still used, and the fact that a subfield of medicine could adopt its own brand of statistical inference in isolation from the rest of the biomedical literature is telling. So, methods education must be both disseminated across fields and targeted within them.

Methods traditions move across scientific disciplines slowly. In psychology, a one-and-done tradition of theory confirmation by a single small randomized trial, while weakening, is still dominant. Multiple articles in the psychology literature over the past decade have raised awareness of the pernicious consequences of single low-power studies, many decades after that became appreciated in clinical medicine. Worse, replication studies in psychology are sometimes taken as an affront to the original researcher (Yong 2012), and the movement to replicate studies has only recently gotten traction.

In the life sciences, the movement has been yet slower, perhaps because the involvement of statisticians in the research is less, and training in statistical reasoning, inadequate. Two of the premier life-science journals, Science and Nature—only added a statistical review component relatively recently (Van Noorden 2014), in contrast to top clinical research journals for which this has been standard since the 1980s. In a poll of life scientists conducted by Nature on ways to improve research reproducibility, the number one remedy named was better training in statistics. (Baker 2016) The leaders of both the NIH and NSF have taken steps and have stated that they are committed to improving research reproducibility in laboratory science.

Gigerenzer and colleagues have written that among disciplines, one finds flavors of statistics, methodologies, standards of proof and acceptable design that are profoundly different (Gigerenzer et al. 1989). Unlike other scientific theories, statistical inference is taught without names or seeming controversy, with the merger of Fisherian and Neyman-Pearsonian approaches called the "silent hybrid solution." (Gigerenzer et al. 1989) Bayesian ideas are rarely mentioned outside of the treatment of Bayesian statistics. It is hard to teach non-frequentist approaches to inference when students are unaware of non-frequentist definitions of probability.

Students who readily accept and understand foundational disputes in physics, biology and economics are treated in statistics courses as though, in Jack Nicholson's immortal cinematic words, they "can't handle the truth." (*https://bit.ly/1sB5MXg*).

Why It Is Hard to Eliminate *P*-Values?

This brings us to the question of why eliminating P-value is so hard. The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It is 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 does not matter if the P-value does not mean what people think it means; it becomes valuable because of what it buys.

Ted Porter, in his book "Trust in Numbers" has written that statistics are a refuge of objectivity for disciplines that are fighting other battles, usually against some form of claimed expertise of its members. Statistics in medicine pushed back against claims of physicians that they could predict what worked and in whom using physiologic or biologic reasoning:

> …the use of statistics tests has become obligatory in…scientific research. …they work mainly as social technologies, not as guides to private thinking.
> The advances of statistics in medicine must be understood as responses to problems of trust, which have been most acute in the context of regulatory and disciplinary confrontations. This, and not any inherently statistical character of clinical medicine, explain why inferential statistics entered medicine through therapeutics. (Porter 1995, p. 209)

P-values are part of a rule-based structure that serves as a bulwark against claims of expertise untethered from empirical support. It can be changed, but we must respect the reason why the statistical procedures are there in the first place. This partly explains why there is so much resistance to Bayesian approaches, which are often viewed as a back-door way to reintroduce the subjectivity that conventional statistical methods were introduced to counter.

The use of *P*-values is a social phenomenon upon which many social rewards and penalties rest. I am faced by this dilemma in my own teaching; when I have biomedical research students for only a day, the first priority is to teach them to be able to read the literature as it is, not as I would like it to be. When I have them for a week, I can include "alternative" material, mainly at the end. I need students for a full term or more, in part to un-teach what they have been taught so I can introduce a different perspective. That's the consequence and power of social norms, which are difficult to change.

A question asked both by Benjamini (2016) and many others in their response to the *P*-value statement is whether *P*-values are really at the heart of the problem. *P*-values carry within themselves conceptual and theoretical baggage that has led to their inappropriate reification and misuse (Goodman 1999a). So it is tempting to think that if we get rid of *P*-values and null hypothesis significance testing, we get rid of that baggage. But that has been the argument for 90 years and thus far is not working.

So what is it that we really want? The ASA statement says it; we want good scientific practice. We want to measure not just the signal properly but its uncertainty, the twin goals of statistics. We want to make knowledge claims that match the strength of the evidence. Will we get that by getting rid of *P*—values? Will eliminating *P*—values improve experimental design? Would it improve measurement? Would it help align the scientific question with those analyses? Will it eliminate bright line thinking? If we were able to get rid of *P*-values, are we sure that unintended consequences wouldn't make things worse? In my idealized world, the answer is yes, and many statisticians believe that. But in the real world, I am less sure.

The problems of *P*-values are in large part from the thoughtless implementation of null hypothesis significance testing (NHST) that even Pearson in his later years said he never envisioned, i.e., divorced from design and from consequences. (Pearson 1962)

The Bayes factor alternative (Goodman 1999b; Kass and Raftery 1995) is attractive but may be the bitcoin equivalent; people are not sure what it means, have little clue where it will be accepted, and it has variations in value. Using Bayes factors or posterior probabilities are akin to forcing people to use the metric system; it makes sense and it fits into a coherent universal system of measurement, but many feel lost when they cannot use familiar measures whose meaning they have internalized, like pounds, Fahrenheit and yards. Bayes factors can also be used in an inappropriate bright-line way if their conceptual foundation as adjusting prior odds are not understood. Bayes factors and methods involve asking for much more than a substitution of one index for another, but rather a new conceptual framework for dealing with scientific uncertainty, which traditional methods do not require probing, and standard courses do not teach how to assess. This is not an argument against some form of Bayes Factor as a replacement for the *P*-value, which I believe would be an improvement (Goodman, 1999b), but an explanation of why its acceptance will be difficult.

The Five "W" Questions

Twenty four of the 31 people who contributed to the ASA *P* value statement development wrote commentaries (Wasserstein and Lazar 2016). That shows how surprising it was that a consensus statement was written at all, akin to a Supreme Court opinion with nine dissents. Of these 24, 21 were generally laudatory, although most were still quite critical of various details. 17 made suggestions for change, 5 being specific (e.g., using a Bayes factor bound in lieu of a *p*-value), and 12 aspirational, w/o a specific strategy (e.g., teach scientists to better understand uncertainty). That showed us that statisticians are passionately interested in the role of statistics in the actual practice of science, but that there is more agreement on what we should *not* do than what we should, a reason why the ASA statement had more "don'ts" then "do's."

In my commentary (Goodman, 2016), I posed the question of what the next steps of statisticians on the road to inferential and research reform are, breaking it down into the 5 "W" questions: who, what, where, when and why. As these are not completely separable, I will address some in combination.

There are a variety of both established and emerging international organizations concerned with improving research conduct, such as the Cochrane and Campbell Collaborations, the Equator Network and the World Congress of Research Integrity. Quite a few centers and initiatives have been established within the last decade to promote reproducible research, such as the Center for Open Science (COS), Berkeley Initiative for Transparency in the Social Sciences (BITSS), and the Meta-Research Innovation Center at Stanford (METRICS). Working with and through any of these entities offer the opportunity to influence the methods and conduct of scientific research locally or worldwide.

Institutions

Some institutional and regulatory doors have already been opened for this, but statisticians must take advantage. A 2018 report released by the National Academy of Sciences entitled "Fostering Integrity in Research" expanded the definition of research integrity beyond the traditional categories of falsification, fraud, and plagiarism to a variety of misleading inferential practices called "detrimental research practices." It suggests that institutions be responsible to monitor and improve both (National Academies of Sciences, 2017) Many if not all of these practices involve deviation from sound statistical design, conduct, analysis and interpretation. Putting proper statistical practice within the frame of scientific integrity, with institutional structures to oversee it may provide an avenue for statisticians to affect the way institutions teach, monitor and promote good science. It means putting structures in place—both training and monitoring—where statistical input can be introduced outside of individual collaborations or standard courses. New teaching models that put inferential issues front and center can also teach and inspire the next generation of both statisticians and scientists to make change.

We also need to create pathways to promotion at our institutions to value and reward faculty who are effective in those activities in addition to than traditional methodological research, coupled with promotion criteria for scientists not based on bibliometrics, which drives the perceived need for significant results (Moher 2018). With the proper professional incentives and rewards, some statisticians might have more time to spend on improving statistical practice, and their collaborators might feel less pressure to game the tools of inference (Wang 2018).

Journals

Applied journals are a very good place for different approaches to be demonstrated, as they can shape the norms of the discipline. Published examples using non-NHST inferential approaches are enormously powerful in demonstrating that papers using these methods will be published in that discipline, as well as illustrating how those methods are to be implemented. Recent publications in JAMA of Bayesian-analyzed clinical trials (Laptook 2017; Goligher 2018) coupled with editorials (Lewis and Angus 2018; Quintana, Viele, and Lewis 2017) serve that role well. Another avenue for influence is for statisticians to serve as statistical editors, of which there are far too few.

Publishing in disciplinary journals also reaches the many teachers and practitioners of statistical methods within disciplines who are not professional statisticians. Writing commentaries that focus on methodology and joining researchers in that field to do their own meta-research, pointing out the prevalence and consequences of suboptimal statistical practice can have great impact, especially if it highlights errors made through practices condemned in the ASA statement.

Funders

Other venue to exert positive influence is with the research funders. It may be that we could do with a little less support for developing new methods and more for tools to train the next generation properly, to produce user friendly code to implement alternatives to *P*-values (*https://osf.io/7dzmk/*), to develop new curricula and hands-on learning tools, and new ways to disseminate and role model proper inferential practice. This is happening to some extent, but the market for such tools is large and constantly evolving.

The most powerful leverage funders have is to require certain levels of methodological rigor in the research they fund. But the transformation of methodological recommendations into implementable policies is difficult. I will present two examples of methods-related policies—one very clear, the other less clear and more problematic in practice. The first is the requirement that any clinical trial submitted to a top medical journal be registered at inception to be considered for publication many years later (DeAngelis et al. 2004). It spurred a dramatic increase in clinical trial registration well before it was required by law in 2007. It was a rule first imposed by the journals, and for the most part, it worked. It was operationally clear, there was a strong technical consensus behind it, the target activity was unambiguous, and the enforcement agent and consequence was spelled out. With the information being public, once this was required by law compliance started to be monitored by others, including the media. (Piller 2015) This preregistration had the effect of increasing the proportion of nonsignificant results in clinical trial publications, and perhaps indirectly decreasing the perceived importance of significant results as a criterion for submission or acceptance (Kaplan, 2015). What the rule and law was less clear about—the completeness and accuracy of the registration—resulted in problems that we are now confronting. (Zarin, 2017).

Another set of methods policies has had only partial success. These are the methodology standards of the Patient-Centered Outcomes Research Institute, PCORI (*https://bit.ly/2xaPcEA*). PCORI was established by Congress with just two boards: the governing board and the methodology committee. This is a surprising arrangement and an opportunity, that a committee of methodologists (currently including a former ASA president) should sit at the highest level of a major research funding agency to advise on how to ensure the best quality science and statistics.

The PCORI methods standards aspire to be minimal, in that they are rules that most would agree if they were violated would seriously weaken or invalidate a given study. They cover how to formulate research questions, the operational meaning of patient centeredness, data integrity, missing data, treatment heterogeneity, data registries, data networks, causal inference, Bayesian trials, diagnostic tests, systematic reviews, data management, and complex interventions.

These are the kinds of standards that one might imagine could prevent the worst problems in scientific practice and inference. But the PCORI effort to translate the standards into policies, assuring that they are met in all research they fund has been only partly successful (Mayo-Wilson, 2017). The reason is that to enforce these standards, a staff member must check a proposal against a long list of multi-part standards that have many grey areas in practice. Even methodologically sophisticated staff find it difficult to judge when the requirements of a standard have been adequately met. If 4% of the data are missing, which cannot be known at the outset, is multiple imputation required? Sensitivity analyses? The consequences of violation are not clear either; should funding be affected by issues like this? This is made yet more complicated when adherence can only be assessed once the project is completed and the data analyzed.

So, when we say that funding agencies have the power to improve the quality of research, whether in technical or inferential dimensions, we have to specify how that power is to be exercised. Often, we find that the entity charged with enforcing change does not have adequate personnel, processes, or enforcement capacity, and that the many gray areas make "violations"—a bright-line designation—hard to assess. The difference between a statistical guideline and a research policy is typically quite wide.

A regulator with a particularly strong influence on statistical practice is the FDA. In addition to issuing guidances on the use of Bayesian methods in device regulation (FDA, 2010), and adaptive trials (including Bayesian approaches) for drugs and biologics, (FDA 2018) the 21st Century Cures Act has directed them to develop further guidances on novel adaptive trial approaches. They are currently funding both internal and external researchers to help advance this effort, and we should expect to see outlines of new FDA research paradigms in the next several years.

The Future

Progress in changing a dominant paradigm is difficult, and lack of movement by any of the key players—funders, journals, institutions, regulators, or payors—constrains movement by the others. The converse is that movement in each of these realms affects the others, and all of these institutions are currently changing. The initial slow speed of progress should not be discouraging; that is how all broad-based social movements move forward and we should be playing the long game. But ball is rolling downhill, the current generation is inspired and impatient to carry this forward, and I trust that by the meetings in 2027, no less 2067, more progress will have been made than we have since 1967. The crisis is upon us, the inference iron is hot, and it is time to get this done.

References

Baker, M. (2016), "1,500 Scientists Lift the Lid on Reproducibility," *Nature*, 533, 452–454. [27]

Batterham A.M., Hopkins W.G. (2006), "Making Meaningful Inferences About Magnitudes," *International Journal of Sports Physiological Performance*, 1, 50–57. [27]

Benjamini, Y. (2016), "It's Not the *P*-values' Fault," *The American Statistician*, ASA Statement supplementary materials, available at *utas_a_1154108_sm5354.pdf* [27]

Butson, M. (2018), "Will the Numbers Really Love You Back: Re-examining Magnitude-Based Inference," [Internet] 2018, available at *https://osf.io/yvj5r/*. [27]

Camerer, C.F., Dreber, A., Holzmeister, F. Ho T-H., Huber J., Johannesson M., Kirchler M., Nave G., Nosek B.A., and Pfeiffer T. (2018), "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015," *Nature Human Behaviour*, 2, 637–644.

Cutler, S. J., Greenhouse, S. W., Cornfield, J., and Schneiderman, M. A. (1966), "The Role of Hypothesis Testing in Clinical Trials," *J Chron Disease*, 19, 857–882. [26]

DeAngelis, C.D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J., Schroeder, T. V., Sox, H. C., and Van Der Weyden, M.B. (2004), "Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors," *JAMA*, 292, 1363–1364. [29]

Echt, D.S., Liebson, P. R., Mitchell, L. B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A., Friedman, L., Greene, H. L., Huther, M. L., Richardson, D. W., and The Cast Investigators (1991), "Mortality and Morbidity in Patients Receiving Encainide, Flecaininde, or Placebo. The Cardiac Arrhythmia Suppression Trial," *The New England Journal of Medicine*, 234, 12 781–788. [26]

Food and Drug Administration, Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials Document (2010), available at *https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf* [29]

Food and Drug Administration (2018), "Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry — Draft Guidance," available at *https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf* [29]

Gigerenzer, G, Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Kruger, L. (1989), *The Empire of Chance*, Cambridge, UK: Cambridge University Press. [27]

Goligher, E.C., Tomlinson, G., Hajage, D. et al. (2018), "Extracorporeal Membrane Oxygenation for Severe Acute Respiratory Distress Syndrome and Posterior Probability of Mortality Benefit in a Post Hoc Bayesian Analysis of a Randomized Clinical Trial," *JAMA*, 320, 2251–2259. [28]

Goodman, S. N. (1999a), "Towards Evidence-based Medical Statistics, I: The P-value Fallacy," *Annals of Internal Medicine*, 130 995–1004. [27]

——— (1999b), "Towards Evidence-based Medical Statistics, II: The Bayes Factor," *Annals of Internal Medicine*, 130 1005–1013. [28]

——— (2016) "The Next Questions: Who, What, When, Where and Why?" *The American Statistician*, 73(S1), this issue. [28]

——— (2018), "How Confident Are You in a Research Finding? Say It with a Number!" *Nature*, 564, 7.

Institute of Medicine (2012), Committee, on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, 'Evolution of Translational Omics'.

Ioannidis, J. P., Fanelli, D., Dunne, D. D. and Goodman, S. N. (2015), Meta-research: Evaluation and Improvement of Research Methods and Practices," *PLoS Biology*, 13 (10), e1002264.

Kaplan, R. M., and Irvin, V. L. (2015), "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time," PLoS One, 10, e0132382. [29]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *JASA*, 90, 773–795. [28]

Laptook, A. R., Shankaran, S., Tyson, J. E., Munoz, B. , Bell, E. F., Goldberg R. N., Parikh, N. A., Ambalavanan, N., Pedroza, C., Pappas, A., Das, A., Chaudhary, A.S., Ehrenkranz, R. A., Hensman, A. M., Van Meurs, K. P., Chalak, L. F., Khan, A. M., Hamrick, S. E. G., Sokol, G. M., Walsh, M. C., Poindexter, B. B., Faix, R. G., Watterberg, K. L., Frantz, I. D., Guillet, R., Devaskar, U., Truog, W. E., Chock, V. Y., Wyckoff, M. H., McGowan, E. C., Carlton, D. P., Harmon, H. M., Brumbaugh, J. E., Cotton, C. M., Sánchez, P. J., Hibbs, A. M., and Higgins, R. D. (2017), "Effect of Therapeutic Hypothermia Initiated After 6 Hours of Age on Death or Disability Among Newborns With Hypoxic-Ischemic Encephalopathy: A Randomized Clinical Trial," *JAMA*, 318 (16), 1550–1560. [28]

Lewis, R. J., and Angus, D. C. (2018), "Time for Clinicians to Embrace Their Inner Bayesian?: Reanalysis of Results of a Clinical Trial of Extracorporeal Membrane Oxygenation," *JAMA*, 320(21), 2208–2210. [28]

Mayo-Wilson, E., Vander Ley, K., Dickersin, K., Helfand, M. (2017), "Patient-Centered Outcomes Research Institute (PCORI) Methodology Standards to Improve the Design and Reporting of Research." [Abstract] International Congress on Peer Review and Publication. *https://peerreviewcongress.org/prc17-0326*. [29]

Micheel C. M. , Nass S. J., Omenn G. S. (2012), *Evolution of Translational Omics : Lessons Learned and the Path Forward*, National Academies Press; Washington, DC. [27]

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., and Goodman, S. N. (2018), "Assessing Scientists for Hiring, Promotion, and Tenure," *PLoS Biol*. 16:e2004089. [28]

National Academies of Sciences, Engineering, and Medicine (2017), "Fostering Integrity in Research," Washington, DC. [28]

PCORI Methodology Committee (2018), "PCORI Methodology Report." available at *https://www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-report*.

Pearson, E.S. (1962), "Some Thoughts on Statistical Inference," *Annals of Mathematical Statistics*, 33, 394–403. [28]

Piller, C. (2015), 'Failure to Report: A STAT Investigation of Clinical Trials', *Stat News*, available at *https://www.statnews.com/2015/12/13/clinical-trials-investigation/* [29]

Porter, T. (1995), *Trust In Numbers: The Pursuit Of Objectivity in Science and Public Life*. Princeton: Princeton University Press. [27]

Quintana, M., Viele, K., and Lewis, R. J. (2017), "Bayesian Analysis: Using Prior Information to Interpret the Results of Clinical Trials," *JAMA*, 318 (16), 1605–1606. [28]

Sainani K. L. (2018), "The Problem with Magnitude-based Inference," *Medicine and Science in Sports and Exercise*, 50 (10), 2166–2176. [27]

Van Noorden, R. (2014), "Science Joins Push to Screen Statistics in Papers," *Nature*, available at *doi:10.1038/nature.2014.15509* [27]

Wang, M.Q., Yan, A. F., and Katz, R. V. (2018), "Researcher Requests for Inappropriate Analysis and Reporting: A U.S. Survey of Consulting Biostatisticians," *Annals of Internal Medicine*, 169 (8), 554–558. [28]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [28]

Yong, E. (2012), "A Failed Replication Draws a Scathing Personal Attack from a Psychology Professor," *Discover Magazine* , available at *https://bit.ly/2qc5bkz*. [27]

Zarin, D. A., Tse, T., Williams, R. J., and Rajakannan, T. (2017), "Update on Trial Registration 11 Years after the ICMJE Policy Was Established," *The New England Journal of Medicine*, 376 (4), 383–391. [29]