Original article

# The role of statistical significance testing in public law and health risk assessment

TOMMASO FILIPPINI[1], SILVIO ROBERTO VINCETI[2]
[1] Environmental, Genetic and Nutritional Epidemiology Research Center (CREAGEN), Section of Public Health,
Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy;
[2] Department of Law, University of Modena and Reggio Emilia, Modena, Italy

Key words

Public law • Statistical significance • Null hypothesis testing • Health risk assessment

## Summary

*Following a fundamental statement made in 2016 by the American Statistical Associations and broad and consistent changes in data analysis and interpretation methodology in public health and other sciences, statistical significance/null hypothesis testing is being increasingly criticized and abandoned in the reporting and interpretation of the results of biomedical research. This shift in favor of a more comprehensive and non-dichotomous approach in the assessment of causal relationships may have a major impact on human health risk assessment. It is interesting to see, however, that authoritative opinions by the Supreme Court of the United States and European regulatory agencies have somehow anticipated this tide of criticism of statistical significance testing, thus providing additional support to its demise. Current methodological evidence further warrants abandoning this approach in both the biomedical and public law contexts, in favor of a more comprehensive and flexible method of assessing the effects of toxicological exposure on human and environmental health.*

## Introduction

Few aspects of scientific methodology as those related to statistical analysis and interpretation, and particularly to statistical significance testing, had and are currently having an effect on causal inference and more generally in the establishment of causal relations in science, including toxicology and biomedical sciences overall but not restricted to them, having major implications also in psychological and economic research [1]. Statistical tests are in fact becoming more complex and sophisticated, frequently relying on an advanced mathematical basis, and are largely employed in medicine and toxicology, among other sciences, to make inferences about causal relations and to inform the risk assessment of interventions such as drugs or of environmental chemicals. Among statistical tests, the most largely used is the so-called "statistically significance testing", based on the evaluation of the compliance of the observed data in any study and experiment with the p-value function and the null hypothesis, i.e. the hypothesis of no association between the chemical or more generally the exposure of interest with the study endpoints [1-4]. In particular, statistical significance testing yields the identification of cut-points based on p-value function, e.g. $p < 0.05$ or $p < 0.001$, subsequently used as reference values for null hypothesis testing, with an ineludible spread of such deleterious and erroneous dichotomous approach relying only on fixed thresholds [5]. Unfortunately, this statistical significance testing has been the pillar and the tenet of risk assessment and biostatistics for decades, despite the unheard complaints by several investigators and methodologists pointing out its ambiguous and confounded information [5, 6]. In somewhat recent times, however, authoritative bodies and scientific communities have raised their voice against the use of p-value and statistical significance testing, invoking the demise of such approach in establishing causation and performing risk assessment [1, 7, 8]. However, the legal world, through pronunciations of the Supreme Court of the United States and scientific contributions by public law scholars, has been advocating the same perspective, i.e., the dismissal of an approach exclusively reliant upon the existence of a dichotomous "statistical significance" in favor of a more flexible and comprehensive method based on a number of factors that include the overall statistical evidence but are not limited to it. We here summarized the history in the use of statistical significance testing and its implication for toxicological risk assessment and for public law, anticipating that the latter will increasingly deal with these methodological issues particularly when dealing with health risks.

## Statistical significance & null hypothesis testing in public health

The statistical training of students and investigators in the biomedical field, including medicine and toxicology, and in other fields such as psychology and economics has been greatly influenced by famous British statistician Ronald Fisher, and more specifically by a small but extremely relevant piece of his intellectual contribution,

https://doi.org/10.15167/2421-4248/jpmh2022.63.1.2394

i.e. the idea of using a single statistical test and even more attractively a single figure to define if results were worth reliance or not in terms of causal inference [1]. Although the influential statistician was not the first to propose the use of p-values, he was the one who suggested a cut-point – 0.05 – "to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach that level" [9]. In other words, Fisher proposed to start from the null hypothesis of being no effect of the investigated "exposure," to compute a p-value function, and to look at the intersection of such function with the effect size observed in the experiment: should such intersection be below 0.05, the results could be considered as "significant" (later considered to mean "statistically significant"). Although Fisher did not encourage to disregard results having a p-value higher than 0.05 and later tempered his position [10], his approach became the boundary line of most scientific inferences based on data analysis in the biomedical and psychological sciences. Results were "significant," i.e., "true" and allowing to reject the null hypothesis of no association, in case p-value was lower than 0.05, further allowing the additional use of the expression "highly significant" in case p-value was < 0.001. By contrast, results exceeding this boundary line were generally dismissed, independently of the actual p-value, and the corresponding results were deemed to be due to chance and not reflecting a causal relation. Unfortunately, such an approach was not accompanied by considerations such as the study sample size (that, if low, inherently increases the p-value for any observed association), the risk of bias of the study, the dose-response relation of the observed phenomena, the biological and temporal plausibility of the associations and finally its consistency across studies, all elements of key relevance when assessing the relation of any cause and exposure to a putative effect as originally suggested by Hill's criteria in 1965 [11], and still relevant when evaluating causal relations in biomedical sciences, especially in public health and toxicology [1, 12]. In many scientific studies and especially in risk assessments, such black and white approach led to the claim that only when p-values are below the 0.05 cut-point we can draw causal inferences and claim the existence of a causal relation between, for instance, a toxic chemical or a drug and any kind of health endpoints.

While many statisticians, methodologists, and even official agencies have long claimed the extreme subjectivity and the serious pitfalls of an approach based on statistical significance and null hypothesis testing, it eventually took almost one century to "officially" highlight these flaws and the most serious implications exerted, for instance, in toxicological risk assessment and in the establishment of causality in legal evaluations. While invitations to consider the fallacious nature in Fisher's claims on statistical significance and p-value cut-points had already been made [1, 13-15], it was only in 2016 that an official statement by the American Statistical Associations officially recognized and highlighted the problem [7]. More recently, a seminal paper that was published in Nature [8] and received the support of a large number of scientists from many disciplines all over the world has convincingly made clear that statistical significance testing and its use in drawing inferences is flawed and may seriously mislead the authors and the readers of scientific articles [2, 3, 16-19]. Along the same line, an increasing number of Editors of scientific journals in the field of epidemiology and public health, medicine, and psychology have accordingly decided to ban or to discourage the reporting of the results as related to "statistical significance testing" [1, 20-23], while putting emphasis on other methodological aspects such as the magnitude and statistical precision of the estimates. Overall, there seems to be an overwhelming majority of methodologists now supporting the demise of statistical significance testing, thus precluding further use of the p-value tool to establish in a black and white manner "causality" in scientific research.

## Recent trends in American public law on the use of statistical significance testing

Contrary to the wide and frequently uncritical propagation of statistical significance testing among scientists in the biomedical field, it is interesting to observe that the legal world has generally been more cautious in its use in scholarly inquiries, as well as in public law practice. This is arguably merit of the long tradition of the legal community in approaching with caution single "absolute" sources of certainty of any type-statistical significance testing undoubtedly and erroneously claiming to be one-and instead weighing the entire body of evidence in favor and against a specific thesis in a more balanced and nuanced way.

A recent example of such a cautious and thoughtful approach, somehow even become a paradigm, can be seen in the 2010 case *Matrixx Initiatives, Inc. v. Syracusan* [24], a seminal decision by United States Supreme Court that has been widely commended and appreciated even beyond the legal circuit [25-29]. The case, involving the pharmaceutical company Matrixx Initiatives, centered on the question of "whether a plaintiff can state a claim for securities fraud based on a pharmaceutical company's failure to disclose reports of adverse events associated with a product" if the reports did not contain statistically significant evidence that the adverse effects may be caused by the use of the product [24]. Delivered by Justice Sonia Sotomayor, the unanimous opinion (9-0) of the Court affirmed the Court of Appeals for the Ninth Circuit's judgment, concluding that the "allegations, 'taken collectively,' give rise to a 'cogent and compelling' inference that Matrixx elected not to disclose the reports of adverse events not because meaningless but because it understood their likely effect on the market 'A reasonable person' would deem the inference that Matrixx acted with deliberate recklessness (or even intent) 'at least as compelling as any opposing inference one could draw from the facts alleged.'. We conclude, in agreement with the Court of Appeals, that respondents have adequately pleaded *scienter*. Whether respondents can ultimately prove their allegations and establish *scienter* is an altogether different question" [24]. The opinion contains several notable statements that directly address the core of the statistical issue at stake, and more generally the basic issues and limitations of statistical significance testing. For instance, the Supreme Court stated that the "lack of statistically significant data does not mean that medical experts have no reliable basis for inferring a causal link between a drug and adverse events" and that "medical experts rely on other evidence to establish an inference of causation." In addition, the Supreme Court emphasized that "medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence." Moreover, "the FDA similarly does not limit the evidence it considers for purposes of assessing causation and taking regulatory action to statistically significant data. In assessing the safety risk posed by a product, the FDA considers factors such as 'strength of the association,' 'temporal relationship of product use and the event,' 'consistency of findings across available data sources,' 'evidence of a dose-response for the effect,' 'biologic plausibility,' 'seriousness of the event relative to the disease being treated,' 'potential to mitigate the risk in the population,' 'feasibility of further study using observational or controlled clinical study designs,' and 'degree of benefit the product provides, including availability of other therapies'". Moreover, the opinion mentions other statements that support the conclusion that statistical significance is not required (and in some cases not achievable) to consider the possibility of causal relations between exposure and an adverse health effect. Overall, the opinion represents an excellent example of correct handling of the concept of statistical significance, under the assumption that it cannot be used as a surrogate indicator of the absence of causal relations. This approach is highly relevant since it goes beyond the traditional approach based on p-value traditional cut-points of 0.05/0.001, dismissing a key role of null hypothesis testing according to Fisher's rule in establishing (and refusing) proof of causation. Unsurprisingly, many scholars have expressed appreciation for this highly relevant opinion, thus indicating how public law theory can take on board a correct approach in dealing with a highly specific and "sophisticated" statistical concept such as statistical significance/null hypothesis testing [25-29]. This comes as no surprise, however, since the issues raised in this seminal sentence by the Supreme Court have long been known to the public law scholarship, as comprehensively illustrated in a relevant paper by David Kaye published as early as 1986 on the Washington Law Review [30].

Most recently, the U.S. Supreme Court has returned to the topic of statistical significance testing in the case *Brnovich v. Democratic National Committee* of March 2021 [31]. Rather than risk assessment and public health, the case dealt with election law and its impact on access to vote. The Democratic National Committee had filed a suit against the State of Arizona's election law since it allegedly "had an adverse and disparate effect on the State's American Indian, Hispanic, and African-American citizens," and had been enacted "with discriminatory intent." To this article, the interesting aspect lies in the statistical significance argument employed by Elena Kagan in her dissenting opinion, where she affirms that Section 2 of the Voting Rights Act of 1965 "demands proof of a statistically significant racial disparity in electoral opportunities" to strike down election rules. Adhering to the Circuit Court's argumentation that voided the District Court's initial dismissal of the suit, Kagan concludes that in the case at hand "Arizona's policy creates a statistically significant disparity between minority and white voters." However, the Court's majority opinion, written by Samuel Alito, rejected what is described as a "procrustean" interpretation of Section 2 of the Voting Rights Act. Citing the Federal Judicial Center's Reference Manual on Scientific Evidence, Alito's majority opinion recalls that "statistical significance may provide 'evidence that something besides random error is at work,' but does not necessarily determine causes." Alito's opinion finds faults with the "statistical manipulation" of emphasizing statistical differences out of a proper context: in that case, while it was factually true that minority voters stood double the chance of having their vote nulled as an out-of-precinct ballot than non-minority voters, the practical difference was in absolute terms so slight that the law could not be held discriminatory.

As a final note, it should be emphasized that not only American public law but also the warnings of European risk assessment institutions signaled and somehow anticipated the shifting tide against the use and misuse of statistical significance testing. For instance, in 2011 the European Food Safety Authority, the official body in charge of assessing the toxicity of food and food constituents, issued a relevant opinion to define how statistical significance testing should (and should not) be used in risk assessment [32]. The opinion represents a good example of the growing awareness, even in a period antecedent to the ASA 2016 statement and the subsequent key scientific contributions, that the dichotomous approach entailed methodological pitfalls and that even in risk assessment null hypothesis testing proved inadequate, despite that being a field generally requiring a final yes/no outcome. The opinion correctly highlighted the need to always report effect/risk estimates and their measures of statistical stability (such as confidences limits), and to give attention to the real biological relevance of the effects even in the presence of small p-values and so-called statistically significant findings [32]. Therefore, it is not surprising that subsequent EFSA assessments and opinions have generally given a limited (if any) reliance on statistical significance testing, putting weight on the strength and the precision of the effect estimates, on dose-response relations, consistency across studies and study designs, quality of the studies and biological plausibility of the associations found in human studies. The convergence in legal and toxicological-epidemiologic approaches toward the rejection of statistical significance testing in risk assessment mirrors the evolution of scientific methodology and appears to be much more adequate to account for all the complexities, the uncertainties but also the potential insights characterizing toxicological risk assessment and its public law implications and litigations [33].

## Conclusions

Implications of abandoning statistical significance testing in public law and to health risk assessment.

For the aforementioned methodological reasons and issues, the approach taken by the U.S. Supreme Court in the case *Matrixx v. Syracusan* case appears to be scientifically sound and somehow even anticipated the methodological shift of several scientific communities, including the statistical one, indicating the growing awareness of the public health community about the pitfalls of simply relying on a conventional black and white approach instead of a balanced assessment of the entire available evidence. Given the large and serious consequences induced by the use of the erroneous approach in data synthesis and causal interpretation represented by statistical significance testing and conventional p-value cut-points, a complete demise of this simplistic approach appears fully justified in both public law and health risk assessment in favor a more challenging but methodologically correct method based on the comprehensive assessment of the strengths and limitations of all the available evidence, and thus abandoning an unwarranted simplification devoid of scientific basis.

## Acknowledgements

None.

## Ethical approval

Ethical approval was not required for this study.

## Conflict of interest statement

The Authors declare no conflict of interest.

## Funding

This study was supported by grant 'UNIMORE FAR IMPULSO 2020' (no. 494/2020) from the University of Modena and Reggio Emilia to Dr. Filippini.

## Authors' contributions

TF and SRV conceived the idea, and equally contributed to the writing of the manuscript.

## References

[1] Berselli N, Filippini T, Adani G, Vinceti M. Dismissing the use of P-values and statistical significance testing in scientific research: new methodological perspectives in toxicology and risk assessment. In: Tsatsakis A, Ed. Toxicological risk assessment and multi-system health impacts from exposure. Elsevier Inc 2021, pp. 309-21. https://doi.org/10.1016/B978-0-323-85215-9.00002-7

[2] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337-50. https://doi.org/10.1007/s10654-016-0149-3

[3] Rothman KJ. Disengaging from statistical significance. Eur J Epidemiol 2016;31:443-4. https://doi.org/10.1007/s10654-016-0158-2

[4] Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Modern Epidemiology. Philadelphia: Wolters Kluwer 2020.

[5] Rothman KJ. Significance questing. Ann Intern Med 1986;105:445-7. https://doi.org/10.7326/0003-4819-105-3-445

[6] Lang JM, Rothman KJ, Cann CI. That confounded p-value. Epidemiol 1998;9:7-8. https://doi.org/10.1097/00001648-199801000-00004

[7] Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016;70:129-33. https://doi.org/10.1080/00031305.2016.1154108

[8] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305-7. https://doi.org/10.1038/d41586-019-00857-9

[9] Fisher RA. The arrangement of field experiments. J Min Agric Great Britain 1926;33:503-13. https://doi.org/10.23637/rothamsted.8v61q

[10] Fisher RA. The design of experiments. London: Oliver and Boyd 1935.

[11] Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965;58:295-300.

[12] Kluxen FM, Jensen SM. Expanding the toxicologist's statistical toolbox: using effect size estimation and dose-response modelling for holistic assessments instead of generic testing. Regul Toxicol Pharmacol 2021;121:104871. https://doi.org/10.1016/j.yrtph.2021.104871

[13] Rothman KJ. A show of confidence. N Engl J Med 1978;299:1362-3. https://doi.org/10.1056/NEJM197812142992410.

[14] Rothman KJ, Greenland S. Modern epidemiology. Second ed. Philadelphia: Lippincott-Raven; 1998.

[15] Nuzzo R. Scientific method: Statistical errors. Nature 2014;506:150-2. https://doi.org/10.1038/506150a

[16] Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. Am J Epidemiol 2017;186:627-35. https://doi.org/10.1093/aje/kwx261

[17] Li G, Walter SD, Thabane L. Shifting the focus away from binary thinking of statistical significance and towards education for key stakeholders: revisiting the debate on whether it's time to de-emphasize or get rid of statistical significance. J Clin Epidemiol 2021;137:104-12. https://doi.org/10.1016/j.jclinepi.2021.03.033

[18] Ciapponi A, Belizan JM, Piaggio G, Yaya S. There is life beyond the statistical significance. Reprod Health 2021;18:80. https://doi.org/10.1186/s12978-021-01131-w

[19] Frank O, Tam CM, Rhee J. Is it time to stop using statistical significance? Aust Prescr 2021;44:16-8. https://doi.org/10.18773/austprescr.2020.074

[20] Trafimow D. Editorial. Basic Appl Soc Psyc 2014;36:1-2. https://doi.org/10.1080/01973533.2014.865505

[21] Lederer DJ, Bell SC, Branson RD, Chalmers JD, Marshall R, Maslove DM, Ost DE, Punjabi NM, Schatz M, Smyth AR, Stewart

PW, Suissa S, Adjei AA, Akdis CA, Azoulay E, Bakker J, Ballas ZK, Bardin PG, Barreiro E, Bellomo R, Bernstein JA, Brusasco V, Buchman TG, Chokroverty S, Collop NA, Crapo JD, Fitzgerald DA, Hale L, Hart N, Herth FJ, Iwashyna TJ, Jenkins G, Kolb M, Marks GB, Mazzone P, Moorman JR, Murphy TM, Noah TL, Reynolds P, Riemann D, Russell RE, Sheikh A, Sotgiu G, Swenson ER, Szczesniak R, Szymusiak R, Teboul JL, Vincent JL. Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. Ann Am Thorac Soc 2019;16:22-8. https://doi.org/10.1513/AnnalsATS.201808-564PS

[22] Harrington D. New guidelines for statistical reporting. Reply. N Engl J Med 2019;381:1597-8. https://doi.org/10.1056/NEJMc1911817

[23] Lin L, Shi L, Chu H, Murad MH. The magnitude of small-study effects in the Cochrane Database of Systematic Reviews: an empirical study of nearly 30,000 meta-analyses. BMJ Evid Based Med 2020;25:27-32. https://doi.org/10.1136/bmjebm-2019-111191

[24] Matrixx Initiatives, Inc., et al., No. 09-1156, Petitioner v. James Siracusano et al., On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit, March 22, 2011, 25 pp., syllabus.

[25] Kaye DH. Trapped in the Matrixx: The U.S. Supreme Court and the need for statistical significance. Prod Saf Liabil Report 2011;39:1007.

[26] Gastwirth JL. Statistical considerations support the Supreme Court's decision. Matrixx Initiatives vs. Siracusano 2012.

[27] Kadane JB. Matrixx v. Siracusano: what do courts mean by 'statistical significance'? Law Probab Risk 2012;11:41-9. https://doi.org/10.1093/lpr/mgr022

[28] Ziliak ST. Statistical significance and scientific misconduct: Improving the style of the published research paper. Rev Soc Econ 2016;74:83-97. https://doi.org/10.1080/00346764.2016.1150730

[29] Ziliak ST, McCloskey D. Lady justice v. cult of statistical significance: oomph-less science and the New Rule of Law. Oxford Handbook of Professional Economic Ethics 2016:352-64. https://doi.org/10.1093/oxfordhb/9780199766635.013.43

[30] Kaye DH. Is proof of statistical significance relevant? Wash L Rev 1986;61:1333-65.

[31] Supreme Court of the United States. Brnovich v. Democratic National Committee, No. 19-1257. 2021:1263.

[32] EFSA Scientific Committee. Scientific opinion: statistical significance and biological relevance. EFSA J 2011;9:2372. https://doi.org/10.2903/j.efsa.2011.2372

[33] Vinceti SR, Filppini T. Towards the dismissal of null hypothesis/statistical significance testing in public health, public law and toxicology. Public Health Toxicol 2021;1:7. https://doi.org/10.18332/pht/144290

Received on November 8, 2021. Accepted on February 10, 2022.

**Correspondence:** Silvio Roberto Vinceti, Department of Law University of Modena and Reggio Emilia, via S. Geminiano 3, 41121 Modena, Italy - Tel.: +39 059 2058170 - Fax: +39 059 2058245 - E-mail: silvioroberto.vinceti@unimore.it

**How to cite this article:** Filippini T, Vinceti SR. The role of statistical significance testing in public law and health risk assessment J Prev Med Hyg 2022;63:E161-E165. https://doi.org/10.15167/2421-4248/jpmh2022.63.1.2394