Towards the dismissal of null hypothesis/statistical significance testing in public health, public law and toxicology

Silvio Roberto Vinceti[1], Tommaso Filippini[2]

AFFILIATION
1 Department of Law, University of Modena and Reggio Emilia, Modena, Italy
2 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

CORRESPONDENCE TO
Tommaso Filippini. 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, Via Campi 287, 41125, Modena, Italy. E-mail: tommaso.filippini@unimore.it ORCID ID: https://orcid.org/0000-0003-2100-0344

KEYWORDS
public law, public health, toxicology, risk assessment, null hypothesis statistical testing

Received: 19 November 2021, **Accepted:** 27 November 2021
Public Health Toxicol 2021;1(2):7
https://doi.org/10.18332/pht/144290

ABSTRACT

Null hypothesis significance testing (NHST) text was once widely popular and almost systematically used for the identification of causal relations and for risk assessment in toxicology and medicine. Interestingly, the public law world has been more prudent and more advanced than the biomedical one in the use of this dichotomous approach, based on the conventional p-value cut-points of 0.05/0.001, to assess causality. The recent 2016 statement by the American Statistical Association, the joint action by methodologists in all fields of science, and not least the seminal decisions by the US Supreme Court have highlighted the pitfalls of the dichotomous approach embedded in NHST. Overall, they also indicated the need to entirely dismiss NHST when assessing causal relations, favoring instead a more flexible and adequate approach for data analysis and interpretation. The demise of statistical significance testing would have major beneficial implications for risk assessment in toxicology, public health, and human medicine, alongside important public law implications. It could also lead to a re-analysis and re-interpretation of previous studies and bodies of evidence that may have been inaccurately assessed due to the flaws inherent in NHST.

INTRODUCTION

Null hypothesis significance testing

Null hypothesis significance testing (NHST) has been, until recently, the standard approach in data analysis and interpretation in most biomedical studies and even beyond this domain, heavily involving other scientific disciplines such as physics, economics, and psychology[1-6]. Indeed, there is probably no other aspect of statistics and methodology, in general, that has affected how data are collected and interpreted within a study, either experimental or observational, either in the human, in animals, or in vitro. This approach goes back to the British statistician Ronald Fisher, at least at the beginning of his intellectual contribution[7,8] but disowned later in his life[9,10]. In summary, the approach assumes that the investigator should compute the so-called p-value function based on the hypothesis that no effect (difference of change in risk) is induced by the variables under study, including toxicological exposure and drugs, and based on the establishment of causal relations on predefined values of such functions, i.e. 0.05/0.001[4-6,11]. This approach was originally formulated by Fisher to reduce the risk of relying too much on 'slight' differences between sets of observations and the risk of attributing effects to risk factors when such effects are actually non-existing. However, this strategy has become so popular as to be considered the 'standard' approach in the large majority of biomedical studies, including the toxicological and risk assessment[1]. Inferences about the existence of causal relations have been

therefore implemented according to these p-value cut-points, i.e. assuming that any observed effect on risk is meaningless not just based on its severity or biological plausibility, or consistency across studies, but just on the overall p-value that could be computed in the studies. Studies in which any effect was 'statistically' characterized by p-values higher than 0.05 were therefore defined as yielding 'non-statistically-significant' findings, or even more shortly and misleadingly non-significant 'findings', i.e. assumed to be irrelevant. This approach is based on the a priori hypothesis that the investigator has to assume that no difference whatsoever exists between exposed and unexposed individuals or organisms (to make it simple and as a general rule) when computation of p-values yields figures above 0.05, as usually occurring in small studies or in studies where effects of the exposure are limited[1,3,4,11]. This is particularly relevant when assessing the effects on safety or on efficacy, i.e. on adverse effects or beneficial ones, and of drugs; therefore basing the final evaluation on a dichotomous, black or white view, using fixed thresholds of the p-value, due to a proposal as a general rule of the statistician who developed it for the first time more than 80 years ago[5]. This also translates into risk assessment, since any assessment made by authoritative bodies (such as the Food and Drug Administration, European Medicine Agency, European Food Safety Authority and any other environmental, nutritional, pharmacological and toxicological agency) or by single investigators, may end up with a final evaluation based on NHST, more than on many other aspects of the assessment. This occurs despite the fact that many methodologists have long highlighted how ambiguous and flawed was the approach, with the possible risk of misleading conclusions, delay in hazard identification, and occurrence of higher detrimental effects for human health and environment[1,10,12-14].

METHODOLOGICAL APPROACH

NHST in toxicology and public law: recent evolutions

NHST definition and suggestion for its use go back to the British statistician Ronald Fisher in 1925, and specifically to his proposal to use a single statistical test (p-value function) and even more a single figure of that parameter (0.05) to define the results of an experiment, or more generally study, as indicative of any effect or causal association[1]. Fisher proposed to use such a p-value cut-point to define the results of the study as really reflecting an effect of the exposure under study, to be named in such a case as 'significant' (later to become 'statistically significant')[8]. Since then, results (i.e. differences in overall measures or proportions, relative risks, odds ratios) characterized by p-values higher than 0.05 have been systematically considered as showing the lack of any association between exposures and effects, either adverse or beneficial. This means that, for instance, toxic factors such as chemicals, and in some cases drugs, that have caused in (human and laboratory) experimental and non-experimental studies adverse effects, have been claimed to be safe and harmless in case these effects occurred accompanied by a p-value higher than 0.05, independently from the strength of the associations (such as the net increase in relative risks or the number of differences, alongside their statistical stability). In such a case, NHST has driven the interpretation of the findings independently from any meaningful interpretation of the effects. Therefore, small studies on toxicological factors and on drugs have generally generated non-statistically significant findings, particularly in the case of rare outcomes, and therefore their results were classified as 'null' and safe in terms of exposure, despite the meaninglessness of such statement. Such misuse of p-values and NHST has therefore led to consider as null many findings of an extremely large number of toxicological studies, despite the interesting clues yielded by these studies that would have been worthy of a much better consideration. On the other hand, the recent shift towards enlarging the number of studies and their size, alongside the implementation of meta-analysis for toxicological and public health assessments, has led to a large increase in the size of the (pooled) study population, thus favoring the generation of 'statistically significant results' even in cases in which the single studies did not individually yield statistically-significant findings.

Only recently, this methodological approach, despite being repeatedly challenged over time[6,15-17], has been definitely acknowledged even at the highest level of the statistical community to be entirely wrong[1,18,19]. In particular, it has been acknowledged that any dichotomization of the p-value or the findings in the two categories of being 'statistically significant' and 'statistically non-significant' is erroneous and must be carefully avoided[1,3,4,10-14,18]. This represents a key change of paradigm versus the previous ongoing approach since Fisher's proposal in 1925, and may have a tremendous and extremely beneficial effect on the reporting and the interpretation of toxicological and public health (and more generally scientific) research, as well as in the quality of publications in scientific journals[20-23]. Adopting this perspective, as a recent Nature article supported by a large consensus from the scientific community has advocated[19], NHST should be quickly expunged from the scientific literature and methodology. In particular, NHST should be abandoned in favor of a considerably more articulated, comprehensive and flexible approach to data analysis and synthesis, taking into consideration the dose-response relation between exposure and toxicological or beneficial effects, the consistency across studies, the biological plausibility of the relations, and the other well-known factors encompassed in the Bradford Hill criteria[15]. This change of perspective may induce some additional complexity in the explanation, reporting and interpretation of the findings of a single study, a meta-analysis or a risk assessment, given the clear simplicity of the black and white NHST approach, but will definitely help in conveying all the relevant findings and the uncertainties embedded in the body of evidence in any assessment[20-24].

Legal wisdom and the stance of the US Supreme Court on NHST

While statistical significance testing has pervaded epidemiology and toxicology in recent decades, the situation in the legal domain remained somehow different, as exemplified by the jurisprudence of the US Supreme Court. 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 for scholarly inquiries, as well as in public law practice. This can arguably be explained by a long tradition of wisdom and prudence in the legal community when approaching allegedly 'unique' 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 or against a specific thesis in a more balanced and prudent way.

A recent example of such a cautious and thoughtful approach, somehow even becoming a paradigm, can be seen in the 2010 case *Matrixx Initiatives, Inc. vs Siracusano*[25], a seminal decision by United States Supreme Court that has been widely commended and appreciated even beyond the legal circuit[2,26-29]. 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[27,30,31]. The unanimous opinion (9 to 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 …' and that '… 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 …' and '… 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'[25].

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 (or refuting) proof of causation. Unsurprisingly, a large number of scholars have expressed their 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[2,26,28,29]. 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 article by Kaye published as early as 1986 on the Washington Law Review[32].

Recently, the U.S. Supreme Court has returned to the topic of statistical significance testing in the case *Brnovich vs Democratic National Committee* of March 2021[33]. 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'. For the purpose of this article, the interesting aspect lies in the statistical significance argument employed in a dissenting opinion, which 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, the dissent concludes that 'Arizona's policy creates a statistically significant disparity between minority and white voters'.

However, the Court's majority opinion, 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, the majority opinion recalls that 'statistical significance may provide evidence that something besides random error is at work, ... but does not necessarily determine causes'. The opinion finds fault with the 'statistical manipulation' of emphasizing statistical differences out of a proper context: in that particular 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.

DISCUSSION

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 (EFSA), 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[34]. The opinion represents a good example of the growing awareness, even in a period antecedent to the American Statistical Association 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 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[34]. 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 strength and precision of the effect estimates, 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.

CONCLUSIONS

Overall, the pattern yielded in the most recent years by scientists in both public health, toxicology, and public law, alongside the US Supreme Court jurisprudence, appear to highlight how incorrect and misleading is NHST in data analysis, interpretation and assessment followed in most instances until recently, to perform safety and efficacy risk assessment and more generally in scientific research. The identification and establishment of causal relations is a complex and difficult endeavor, that requires time, considerable effort, and unavoidably entails some subjectivity (that must be clearly and honestly acknowledged and reported) and cannot undergo oversimplification such as that represented by NHST. The conventional 'black and white' approach must be therefore avoided whenever the identification of causal links is pursued, both within and outside public health, toxicology and public law, and both the public and the professionals must be aware of the pitfalls of simplification based on conventional p-value cut-points and NHST. This also emphasizes the relevance of adequate statistical training and reporting to avoid overconfidence about the potential of some conventional statistical approach in data interpretation and in risk assessment, and eventually of using the full spectrum of tools and evidence to assess and establish (or deny) causal relations, starting from the Bradford-Hill criteria.

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 AM, ed. Toxicological Risk Assessment and Multi-System Health Impacts from Exposure. Elsevier; 2021:309-321. doi:10.1016/B978-0-323-85215-9.00002-7
2. Ziliak ST. Statistical significance and scientific misconduct: improving the style of the published research paper. Rev Soc Econ. 2016;74(1):83-97. doi:10.1080/00346764.2016.1150730
3. Rothman KJ, Lash TL, VanderWeele TJ, Haneuse S. Modern Epidemiology. 4th ed. Wolters Kluwer; 2021.
4. Rothman KJ. Disengaging from statistical significance. Eur J Epidemiol. 2016;31(5):443-444. doi:10.1007/s10654-016-0158-2
5. Rothman KJ. Significance questing. Ann Intern Med. 1986;105(3):445-447. doi:10.7326/0003-4819-105-3-445
6. Rothman KJ. A show of confidence. N Engl J Med. 1978;299(24):1362-1363. doi:10.1056/NEJM197812142992410
7. Fisher RA. The design of experiments. Oliver and Boyd; 1935. Accessed November 19, 2021. http://tankona.free.fr/fisher1935.pdf
8. Fisher RA. The arrangement of field experiments. Journal of the Ministry of Agriculture. 1926;33:503-513. doi:10.23637/rothamsted.8v61q
9. Gigerenzer G. Mindless statistics. Journal of Socio-Economics. 2004;33(5):587-606. doi:10.1016/j.socec.2004.09.033
10. 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. doi:10.1016/j.yrtph.2021.104871
11. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3
12. 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;137:104-112. doi:10.1016/j.jclinepi.2021.03.033
13. Ciapponi A, Belizán JM, Piaggio G, Yaya S. There is life beyond the statistical significance. Reprod Health. 2021;18(1):80. doi:10.1186/s12978-021-01131-w
14. Frank O, Tam CM, Rhee J. Is it time to stop using statistical significance? Aust Prescr. 2021;44(1):16-18. doi:10.18773/austprescr.2020.074
15. Hill AB. The Environment and Disease: Association or Causation? Proc R Soc Med. 1965;58(5):295-300. doi:10.1177/003591576505800503
16. Rothman KJ, Greenland S. Modern Epidemiology. 2nd ed. Lippincott Williams;1998.
17. Lang JM, Rothman KJ, Cann CI. That confounded P-value. Epidemiology. 1998;9(1):7-8. doi:10.1097/00001648-199801000-00004
18. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108
19. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305-307. doi:10.1038/d41586-019-00857-9
20. Harrington D. New Guidelines for Statistical Reporting. Reply. N Engl J Med. 2019;381(16):1597-1598. doi:10.1056/NEJMc1911817
21. Lederer DJ, Bell SC, Branson RD, et al. 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(1):22-28. doi:10.1513/AnnalsATS.201808-564PS
22. Trafimow D. Editorial. Basic and Applied Social Psychology. 2014;36(1):1-2. doi:10.1080/01973533.2014.865505
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(1):27-32. doi:10.1136/bmjebm-2019-111191
24. Hartung T, Tsatsakis AM. The state of the scientific revolution in toxicology. ALTEX. 2021;38(3):379-386. doi:10.14573/altex.2106101
25. Matrixx Initiatives Inc vs Siracusano, 563 US 27 (2011). Accessed November 19, 2021. https://supreme.justia.com/cases/federal/us/563/27/
26. Kaye DH. Trapped in the Matrixx: The U.S. Supreme Court And the Need for Statistical Significance. bloomberglaw.com. November 9, 2011. Accessed November 19, 2021. https://news.bloomberglaw.com/product-liability-and-toxics-law/trapped-in-the-matrixx-the-us-supreme-court-and-the-need-for-statistical-significance
27. Gastwirth JL. Statistical Considerations Support the Supreme Court's Decision in Matrixx Initiatives V. Siracusano. Jurimetrics. 2012;52:155-175. doi:10.2139/ssrn.1925465
28. Kadane JB, Matrixx v. Siracusano: what do courts mean by 'statistical significance'? Law, Probability and Risk, 2012;11(1):41-49. doi:10.1093/lpr/mgr022
29. Ziliak ST, McCloskey D. Lady Justice Versus Cult of Statistical Significance: Oomph-less Science and the New Rule of Law. In: George F. DeMartino GF, McCloskey D, eds. The Oxford Handbook of Professional Economic Ethics. Oxford University Press; 2016:352-364. doi:10.1093/oxfordhb/9780199766635.013.43
30. Shook B. The Materiality Standard after Matrixx Initiatives,Inc. v. Siracusano. N C J Law Technol. 2011;12(2):369-384. Accessed November 19, 2021. http://scholarship.law.unc.edu/ncjolt/vol12/iss2/6
31. Leisawitz BA. Matrixx Initiatives, Inc v Siracusano: Rejection of the Statistically Significant Standard Reopened the Door to Securities Fraud Strike Suits. Delaware Journal of Corporate Law. 2011;36(2):675-706. Accessed November 19, 2021. https://ssrn.com/abstract=1950417
32. Kaye DH. Is Proof of Statistical Significance Relevant? Wash Law Rev. 1986;61(4):1333-1365. Accessed November 19, 2021. https://elibrary.law.psu.edu/fac_works/52/
33. Brnovich v Democratic National Committee, 594 US (2021). Accessed November 19, 2021. https://supreme.justia.com/cases/federal/us/594/19-1257/
34. EFSA Scientific Committee. Statistical Significance and Biological Relevance. EFSA J. 2011;9(9):2372. doi:10.2903/j.efsa.2011.2372

CONFLICTS OF INTEREST
The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none was reported.

FUNDING
There was no source of funding for this research.

ETHICAL APPROVAL AND INFORMED CONSENT
Ethical approval and informed consent were not required for this study.

DATA AVAILABILITY
Data sharing is not applicable to this article as no new data were created.

PROVENANCE AND PEER REVIEW
Not commissioned; externally peer reviewed.