Statistical significance abuses in public health research: a retrospective investigation of recent trends and possible solutions

Alessandro Rovetta, R&C Research, Bovezzo (BS), 25073, Italy

Correspondence

Email: rovetta.mresearch@gmail.com

Phone: +39 3927112808

ORCID: https://orcid.org/0000-0002-4634-279X

Abstract

Background. Despite the efforts of leading statistical authorities and experts worldwide and the inherent dangers of interpretative errors in clinical research, misuses of statistical significance remain a common practice in the field of public health. Currently, there is a need to attempt to quantify this phenomenon.

Methods. 100 studies were randomly selected within the PubMed database. An evaluation system for the interpretation, presentation, and communication of results (IPC) was adopted, which provided for a maximum of 11 points and a minimum acceptability threshold of 5 points.

Results. The median of the results was 2 points out of the available 11 (IQR = 1). The difference from the minimum acceptable IPC score of 5 was substantial (90|95|99-% CI: [2; 4]) and, assuming all the Wilcoxon test requirements have been sufficiently met, highly surprising at the statistical level (P-value < .001, S-value < 10). In total, 13 out of 100 studies achieved the minimum score of 5 points.

Conclusion. These findings provide solid evidence of widespread and severe methodological shortcomings in the use of statistical significance measures in clinical and public health research during 2023. Therefore, it is essential for academic journals to compulsorily demand higher scientific quality standards.

Keywords. Clinical research; infodemic; public health; scientific publishing; significance fallacy; statistical significance.

Introduction

Despite decades of intense informative efforts, the misuse of the concept of statistical significance remains one of the primary and most pervasive issues within the scientific community, particularly in the field of public health [1]. In general, there is a prevailing tendency to interpret the p-value as an objective measure capable of discerning scientifically significant results from those that are not [2]. However, as pointed out by its own creator, Ronald Fisher, and reiterated by numerous experts, including Sander Greenland, such a practice is entirely unfounded and contradicts the most recent evidence on the topic [3, 4]. Indeed, when used appropriately, the p-value can assist in assessing the statistical relevance of results, but it remains a measure subject to very high and ineliminable margins of uncertainty. Furthermore, authors frequently blur the lines between the statistical surprise of a result and its clinical relevance, although these being two entirely separate aspects [5]. However, the purpose of this paper is not to reiterate concepts extensively discussed in the literature, but rather to conduct an investigation aimed at estimating recent trends in the adoption of these measures. Given that these kinds of errors can have severe consequences in the healthcare sector, including the approval of ineffective drugs or the rejection of effective ones, such an evaluation is both a priority and an urgent necessity. It should be clarified that the methodologies used to assess statistical significance in this paper adhere to the guidelines set forth by the American Statistical Association [6].

Methods

Selection criteria and collection procedure

The PubMed database of the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) was consulted for the study as it represents one of the most important repositories of scientific peer-reviewed medical articles in the world. To have a representative sample of the most recent trends, only the year 2023 was selected (as of 3 September 2023). The search keyword was “public health AND regression” since regression represents one of the main methods of quantitative investigation in the field of public health. In addition, this increased the likelihood of finding studies containing analyses based on statistical significance. The search returned about 6011 results. Through a random generator of integers from 1 to 6011 with a uniform probability distribution, 100 studies with the following characteristics were selected: i) the study concerned public health topics, 2) the study had an open access abstract, and 3) the latter contained quantitative results in which the statistical significance of the results was evaluated. The methodology section of the full papers was often consulted to verify the approach adopted regarding statistical significance.

Evaluation process

To evaluate the quality of the presentation of the results, six categories were defined according to the scheme shown below.

1. Significance continuity. The purpose was to evaluate whether the P-value was treated as a continuous measure (1 point) or not (0 points). In this regard, the possible use of dichotomous expressions “significant/non-significant” or thresholds (e.g., α=.05) was considered.

2. Full p-values. The purpose was to evaluate whether the P-values were reported in full for all results (2 points), only for results considered statistically significant (1 point), or never reported in full (0 points).

3. Effect size measures. The purpose was to evaluate the adoption of effect size measures for all results (2 points), only for results considered statistically significant (1 point), or never reported (0

points). Confidence/compatibility intervals, standard errors, or specific measures such as Cohen's D were considered effect size measures.

4. Effect size comments. The purpose was to evaluate the adoption of effect size ranges (e.g., the distinction between small, medium, and large effects) for all results (2 points), only for results considered statistically significant (1 point), or for no results (0 points).

5. Best estimators. The purpose was to evaluate the adoption of the best effect estimators (e.g., correlation and regression coefficients) for all results (2 points), only for results considered statistically significant (1 point), or for no results (0 points).

6. Proper language. The purpose was to evaluate the tone of the considerations built on the basis of the results found. The tone was considered appropriate when only expressions such as "this evidence suggests that" (2 points) were adopted, sensationalistic when expressions such as "these results demonstrate" or when P>/<α was confused with the absence/presence of clinical significance or effects (0 points) were adopted, and acceptable when mixed expressions such as "these findings prove that […] However, further studies are needed to fully validate them." (1 point) were adopted.

The maximum score obtainable was therefore 11. Nevertheless, since it is not always possible to perfectly summarize the results obtained in the abstract due to purely logistical reasons and there is a component of subjectivity in the assessment of the sixth category, a score of 5 was considered as the threshold of acceptability. It is essential to specify that this paper does not investigate the methodological rigor with which the studies were conducted, but rather focuses solely on the interpretation, presentation, and communication (IPC) of the results obtained.

Statistical analysis

The Wilcoxon signed-rank test was used to assess the statistical surprise in the difference between the expected minimum score (5 points) and the median of the distribution obtained from the examination of the 100 studies. The effect size was assessed by adopting multiple confidence intervals (90%, 95%, and 99%), calculated using the bootstrap method with 1000 repetitions with the software R-studio (v.4.2.0). The inspection of the frequency histogram revealed sufficient symmetry.

Results

The median of the results was 2 points out of the available 11 (IQR = 1). The difference from the minimum acceptable IPC score of 5 was substantial (90|95|99-% CI: [2; 4]) and, assuming all the Wilcoxon test requirements have been sufficiently met, highly surprising at the statistical level (P-value < .001, S-value < 10). In total, 13 out of 100 studies achieved the minimum score of 5 points. **Table 1** reports the results by category. As can be observed, statistical significance was treated as a continuous measure in 3 out of 100 cases. Most studies reported p-values, estimators, and effect size measures for results with P < α but not for those with P > α; 14 studies presented only p-values without displaying any effect size estimators. In most cases, sensational language was adopted, and no comments on the effect size were explicitly reported.

	Significance continuity	*Full p-values*	*Effect size measures*	*Effect size comments*	*Best estimators*	*Proper language*
≥ 1 point	3	52	57	11	86	42
= 2 points	N.A.	8	1	0	3	7

Table 1. This table shows the percentage of studies (n=100) that passed 1 point or achieved the maximum of 2 points in each assessment category.

Discussion

This survey found that the median IPC score of the 100 randomly selected medical articles of 2023 was markedly lower than the minimum acceptable score. Indeed, only a small proportion of papers reached the required scientific quality in disclosing the outcomes. When analyzing the results by category, it was observed that a very limited number of studies treated statistical significance as a continuous measure. Moreover, the vast majority of them reported statistical measures only for results whose p-value fell below the previously established threshold, and some have even reported only measures of statistical significance without any effect size estimators. Such a scenario is wholly unsatisfactory, particularly when placed within the context of public health and safety. While governmental and health agencies such as the World Health Organization, Centers for Disease Control and Prevention, Food and Drug Administration, and European Medicines Agency have their internal evaluation committees dedicated to ensuring the clinical efficacy of treatments and drugs, these widespread errors and uncertainties in the field of clinical research can not only propagate a marked infodemic – as often witnessed during the COVID-19 pandemic – but also result in a wastage of resources, such as the prolonged funding for studies with exaggerated outcomes [7-9]. In fact, sensationalistic expressions are aimed at increasing the perception of the study's relevance beyond its actual findings, i.e., to boost the number of citations and success, a crucial factor in securing research funding and even institutional roles. Given that, as highlighted by the undersigned and various experts in the field as well as supported by these findings, there is a furious resistance to changing these scientifically unsound practices, the author of this manuscript calls for academic journals to begin mandating scientific standards that align with the latest statistical evidence advocated by organizations such as the American Statistical Association. Furthermore, journal editorial policies should assign equal weight to both positive and negative findings. This must be done in the name of scientific and medical ethics since it is an essential step toward conducting unbiased investigations. Based on this, the following basic recommendations are proposed. First, only if all test assumptions are sufficiently met (a methodological aspect to be extensively discussed in the manuscript, especially when dealing with clinical results), academic journals should explicitly and compulsorily require that p-values be treated as a continuous measure of the compatibility between the test result and the target hypothesis (e.g., null hypothesis). Specifically, p-values close to 1 indicate high compatibility, while p-values close to 0 indicate low compatibility. Second, academic journals should explicitly and compulsorily require that the effect size be treated as a completely separate aspect from statistical significance. Third, academic journals should explicitly and compulsorily require that authors refrain from using sensationalistic expressions when presenting results, especially if the latter stem from statistical analyses. As a matter of fact, statistics can provide – when hypotheses are well-targeted, i.e., motivated by evidence of other kinds – further evidence in favor of or against a phenomenon but can never, in any way, prove or disprove its existence. In this regard, it must be emphasized that the p-value refers only to the test result and not the phenomenon under investigation in itself. Finally, academic journals should explicitly and compulsorily require that scientific recommendations must be provided on the basis of an analysis of the risks, costs and benefits and not on the p-value or any other statistical indicator [10].

Conclusion

These findings provide solid evidence of widespread and severe methodological shortcomings in the use of statistical significance measures in clinical and public health research during 2023. This is consistent with decades of criticism from epidemiologists and statisticians, including respected international organizations. Such errors can result in highly misleading interpretations, posing a threat to public safety. As a result, it is essential for academic journals to demand higher scientific quality standards.

Ethical Considerations

The author declares that he has no conflicts of interest.

Funding

No funding was obtained for this research.

References

1. 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 Apr;31(4):337-50. doi: 10.1007/s10654-016-0149-3. Epub 2016 May 21. PMID: 27209009; PMCID: PMC4877414.

2. Rovetta A. A Framework to Avoid Significance Fallacy. Cureus. 2023 Jun 11;15(6):e40242. doi: 10.7759/cureus.40242. PMID: 37440801; PMCID: PMC10334213.

3. Connecting simple and precise P-values to complex and ambiguous realities (includes rejoinder to comments on “Divergence vs. decision P-values”) Greenland S. Scand J Stat. 2023:1–16.

4. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ. 2017 Jul 7;5:e3544. doi: 10.7717/peerj.3544. PMID: 28698825; PMCID: PMC5502092.

5. Kühberger A, Fritz A, Lermer E, Scherndl T. The significance fallacy in inferential statistics. BMC Res Notes. 2015 Mar 17;8:84. doi: 10.1186/s13104-015-1020-4. PMID: 25888971; PMCID: PMC4377068.

6. Yaddanapudi LN. The American Statistical Association statement on P-values explained. J Anaesthesiol Clin Pharmacol. 2016 Oct-Dec;32(4):421-423. doi: 10.4103/0970-9185.194772. PMID: 28096569; PMCID: PMC5187603.

7. Ogbodo JN, Onwe EC, Chukwu J, Nwasum CJ, Nwakpu ES, Nwankwo SU, Nwamini S, Elem S, Iroabuchi Ogbaeja N. Communicating health crisis: a content analysis of global media framing of COVID-19. Health Promot Perspect. 2020 Jul 12;10(3):257-269. doi: 10.34172/hpp.2020.40. PMID: 32802763; PMCID: PMC7420175.

8. Rovetta A. Health communication is an epidemiological determinant: Public health implications for COVID-19 and future crises management. Health Promot Perspect. 2022 Dec 10;12(3):226-228. doi: 10.34172/hpp.2022.28. PMID: 36686052; PMCID: PMC9808906.

9. Saracco A. Dr. Strangelove: Or How I Learned to Stop Worrying and Love the Citations. Math Intelligencer. 2022;44:326-330. doi:10.1007/s00283-021-10146-x.

10. Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. Paediatr Perinat Epidemiol. 2021 Jan;35(1):8-23. doi: 10.1111/ppe.12711. Epub 2020 Dec 2. PMID: 33269490.