#01/2024
ISSN 2299-1247

JHPOR | Journal of Health Policy & Outcomes Research

Statistical significance misuse in public health research: an investigation of the current situation and possible solutions

DOI:10.7365/JHPOR.2024.1.7

Authors:

Alessandro Rovetta
orcid.org/0000-0002-4634-279X

R&C Research

Keywords:

Clinical research; infodemic; public health; significance fallacy; statistical significance.

How to cite this article?
Rovetta A., Statistical significance misuse in public health research: an investigation of the current situation and possible solutions J Health Policy Outcomes Res [Internet]. 2024[cited YYYY Mon DD];. Available from: https://jhpor.com/article/2357-statistical-significance-misuse-in-public-health-research-an-investigation-of-the-current-situation-and-possible-solutions

contributed: 2023-11-07
final review: 2024-04-22
published: 2024-05-07
Corresponding author:
Alessandro Rovetta rovetta.mresearch@gmail.com

Abstract

Background
Despite the efforts of leading statistical authorities and experts worldwide, misuse of statistical significance remains a common, dangerous practice in public health research. There is an urgent need to quantify this phenomenon.

Methods
200 studies were randomly selected within the PubMed database. An evaluation scale for the interpretation and presentation of statistical results (SRPS) was adopted. The maximum achievable score was 4 points. Abstracts (A) and full texts (FT) were compared to highlight any differences in presentation. The Wilcoxon signed-rank test was employed in this scope.

Results
All studies failed to adopt P-values as continuous measures of compatibility between the data and the target hypothesis as assessed by the chosen test. The vast majority did not provide information on the model specification. However, in most cases, all findings were reported in full within the manuscripts. The Wilcoxon signed-rank test showed a marked incompatibility of the null hypothesis of zero difference between A and FT scores with the data obtained in the survey: null P < .001 (as assessed by the model), r = 0.87 (standardized effect size). Additionally, the score difference (207.5 points for A vs. 441.5 points for FT) indicates a scenario consistent with a substantial disparity in the completeness of the outcomes reporting.

Conclusion
These findings align with the hypothesis of widespread and severe shortcomings in the use and interpretation of statistical significance within public health research during 2023. Therefore, it is essential for academic journals to compulsorily demand higher scientific quality standards. The suggestions provided in this study could be helpful for this purpose.

Introduction

Background
Decades of intense efforts had little impact on the misuse of the concept of statistical significance, which 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 for discerning between scientifically significant and non-significant results.[2] However, such a practice is entirely unfounded and contradicts consolidated evidence on the topic.[1-6] There are two main, mutually exclusive approaches within the frequentist scenario: the neo-Fisherian and the Wald-Neyman-Pearson (WNP) ones.[7] And, paradoxically, despite the historical diatribe between these authors and the extreme philosophical and epistemological differences, such approaches are erroneously mixed in much of today's research. In the former, at best, the so-called "divergence P-value" is a continuous measure of the statistical compatibility between the numerical, experimental data and the fixed target hypothesis (e.g., the null hypothesis of no effect) as evaluated by the chosen statistical model.[8] This measure is conditional on the background assumptions (e.g., absence of bias, extraction of a random and sufficiently large sample from the target population, numerical data normality, etc.), which must be well met for the P-value to be sufficiently interpretable in that way (the higher the P-value, the greater the statistical compatibility). The term "divergence" stems from the fact that such a P-value can also be adopted to assess the discrepancy of a statistical result from the prediction of the target hypothesis (the lower the P-value, the greater the discrepancy). This discrepancy, or incompatibility, is generally and improperly referred to as "significance." On the contrary, the so-called "decision P-value" of WNP is a mere numerical index devoid of intrinsic, direct meaning: given a certain threshold α (e.g., .05), one arbitrarily chooses to reject the target hypothesis in favor of the alternative hypothesis when P-value < α and not to reject it when P-value ≥ α (according to what is computed by the chosen test).[8] A little-understood but yet vital aspect of this procedure is that, in addition to being conditional on various underlying hypotheses, it is mathematically structured to work only in a high number of ideal experimental executions (each capable of sufficiently guaranteeing the required underlying hypotheses) and is mathematically precluded from providing specific information on the individual study already conducted. The modern utopian goal is to limit the number of false positives - or Type I errors - to a maximum of α% (and, eventually, false negatives - or Type II errors - to a maximum of β%) in the set of all experiments, without being able to determine in which of the individual experiments a wrong decision was made.[9, 10] As noted by Fisher himself and

reiterated by leading global authorities in the field (including the American Statistical Association), in complex scientific contexts - due to sources of uncertainty difficult to manage and even identify - the WNP approach is unjustified (it is not structured to manage non-equivalent replications, Type III errors).[5, 11] The mathematical and epistemological impossibility of informing decisions on individual studies also makes it potentially misleading for public health.[8, 10-12] Moreover, the presumed objectivity of the model is inseparably linked to an ineliminable degree of subjectivity in its selection and the evaluation of the required background assumptions (specification).[12] In this regard, as emphasized by Professor Sander Greenland in a recent seminar at Harvard, cognitive distortions, moral values, personal beliefs, and other personal characteristics also act as sources of uncertainty within scientific and consequently statistical investigation.[12-14] Therefore, from now on, we will refer to the P-value as the neo-Fisherian divergence P-value, emphasizing the need for an (almost) unconditional/descriptive approach in its interpretation (i.e., assessing and exposing the compatibility of numerical data with all relevant scientific hypotheses equally, including the recognized limitations). [15]

Common errors in statistical testing

There are further widespread misinterpretations of statistical testing. The so-called "nullism," namely, the tendency to consider only the mathematically null hypothesis of no effect (e.g., hazard ratio HR = 1.00) never allows for a complete picture of real clinical significance.[16] Indeed, situations may arise where the P-value for the mathematically null hypothesis HR = 1.00 indicates, according to the chosen methods and assuming the above ideal conditions, marginal compatibility with numerical data (e.g., P = 0.03), and at the same time, the P-value for the mathematically non-null but practically weak (in many clinical settings) hypothesis HR = 1.05 indicates perfect compatibility with numerical data (e.g., P = 1, best estimate). Cases like this, generally labeled as "significant" (since P < .05 for the sole null hypothesis), don't prove nor suggest the occurrence of a significant phenomenon but rather, at most, show a statistical result consistent with a non exactly null but still small effect at the clinical/epidemiological level. But not only that: a 90% "confidence interval” (which, from now on, we will call “compatibility interval” for the reasons mentioned above) of the form (0.99, 15.0) signals, conditionally on the background assumptions, only a large statistical uncertainty and not the absence of statistical or any other significance.[17] Indeed, the hypothesis of an almost null effect size HR = 0.99 and the hypothesis of a very large effect size HR = 15.0 are equally compatible with numerical data (P-value = .10, as they are the 90% compatibility interval limits). This is why it is particularly important to test the compatibility of numerical data with various hypotheses of scientific interest or to show, at least, a compatibility interval alongside the sole P-value for the mathematically null hypothesis. For instance, suppose we obtain HR = 3.2, 95% CI (1.0, 10.0). Many would label this outcome as non-significant because the mathematically null hypothesis is contained within the 95% compatibility interval (P-value = .05). However, the hypothesis most compatible with the data is certainly not the mathematically null one but rather the decidedly non-null best estimate HR = 3.2 (P-value = 1). Finally, it is worth reiterating that frequentist statistics is neither mathematically nor epistemologically structured to support scientific hypotheses (e.g., the functioning of a drug), and it is always the combination of various, concordant pieces of evidence (e.g., biochemical, clinical, psychological, physical, etc.) that can provide initial indications of causal mechanisms.[18-20] Given that the current (mis)use of statistical significance can have severe consequences in the healthcare sector, including the approval of ineffective drugs or the rejection of effective ones, it is essential to evaluate recent trends on this matter.[17-20]

Context and objectives

To ensure maximum transparency in data interpretation, the author briefly describes the motivations – and, consequently, the potential biases – that led him to undertake this research. The primary motivation is the author's personal experience with these errors in the past and his desire to contribute to preventing them in future studies. The secondary motivation arises from his roles as a peer reviewer and editor, where he has encountered this type of misuse in over 80% of the manuscripts analyzed. The research objective is to quantify the phenomenon in the current context as of October 2023. Specifically, this does not involve assessing the overall quality of the manuscripts, not even from a purely statistical perspective, but solely focuses on the use of neo-Fisherian incompatibility (or, improperly, significance) to inform public health decisions (also distinguishing between statistical incompatibility and statistical effect size). It is also clarified that the aim is not to obtain a very precise estimate but rather a preliminary indication of prevalence. Furthermore, the author declares that he is not interested in the potential role of journals in the above mistakes (e.g., editorial requirements about P-values) but only in the overall infodemic scenario. Based on these considerations and the literature he has read on the subject, the author hypothesizes that a considerable number of misuses will be observed (potential bias).

Materials and 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, the current month was selected (from 11 September to 11 October, 2023). The search keyword was: “ANOVA” OR “regression” OR “t-test” OR “Chi square” OR “Mann-Whitney U test” OR “Kruskal-Wallis test” OR “Fisher’s exact test” OR “Logrank test” OR “Kolmogorov–Smirnov test” OR “Wilcoxon signed-rank test” OR “Dunnett’s test” OR “ANCOVA” OR “Levene’s test” OR “Friedman test” OR “Pearson correlation” OR “Spearman correlation” OR “Kendall correlation”. This choice was made considering the most commonly used statistical methods in the field of public health.[7-10] In addition, this increased the likelihood of finding studies containing analyses based on statistical significance. The search returned about 7,771 results. Through a random generator of integers from 1 to 7,771 with a uniform probability distribution, 200 studies with the following characteristics were selected: i) the study concerned public health topics, 2) the study contained quantitative results both in the text and the abstract, and 3) the study was peer-reviewed. The general process is summarized in Figure 1.

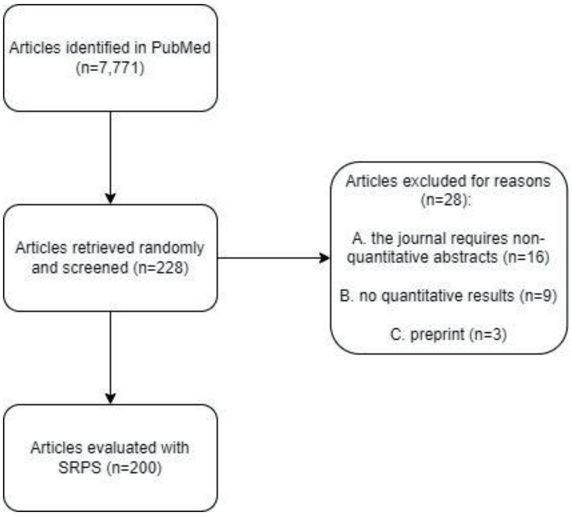
Articles identified in PubMed (n=7,771)
Articles retrieved randomly and screened (n=228)
Articles excluded for reasons (n=28):
A. the journal requires non-quantitative abstracts (n=16)
B. no quantitative results (n=9)
C. preprint (n=3)
Articles evaluated with SRPS (n=200)

Figure 1. Articles selection and collection procedure flow chart.

Evaluation scale

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

1. Incompatibility (significance) continuity for the target hypothesis. 1 point is awarded if and only if statistical incompatibility is measured continuously, i.e., the P-value is used as a continuous index of the data incompatibility with the target hypothesis as conditionally assessed by the chosen statistical test. In all other cases, i.e., when a threshold is adopted and/or when some results are referred to as “non-significant” and others as “significant,” 0 points are awarded.
2. Full P-values for the target hypothesis (called "null P" if solely referred to the null hypothesis). 1 point is awarded if and only if all P-values are reported in full (unless P < .001) for all tests. 0.5 points are awarded when P-values are reported in full (unless P < .001) only for some measures (e.g., those considered significant); this includes mixed cases like P = .02 and P > .05. 0 points are awarded when no P-values are reported in full. It is specified that notations like P < .05 fall into the latter situation.
3. Global effect size measures. Compatibility/confidence intervals or distributions, standard errors, and specific measures like Cohen's D or Hedges' g have been included in this category. 1 point is awarded if and only if measures of statistical effect size have been reported for all conducted tests. 0.5 points have been awarded if measures of statistical effect size are reported for some of the tests conducted or for individual measures before testing (e.g., reporting two means with respective standard errors but not a compatibility interval of the difference after the z-test). 0 points are awarded when measures of statistical effect size are not reported for any test.
4. Best estimates. 1 point is awarded if and only if the best estimates (e.g., correlation/regression coefficients, percentage differences, odd ratios, etc.) are reported for all conducted tests. 0.5 points are awarded if the best estimates are reported for some of the conducted tests (e.g., those considered "significant"). 0 points are awarded when measures of statistical effect size have not been reported for any test.
5. Background assumptions (full text only). 1 point is awarded if and only if a complete assessment of the background assumptions of the tests adopted is described and reported in the paper or supplementary materials. 0.5 points are awarded if a complete assessment of the background assumptions is described but not reported. 0 points are awarded in all other cases.

This scale, called SRPS (statistical results presentation scale), was designed by considering the basic elements for a comprehensive evaluation of a statistical effect in a single study. These include the use of the neo-Fisherian approach (category 1), the reader's ability to assess the conditional compatibility of the null hypothesis (category 2), and the distinction between the size of the statistical effect and the conditional compatibility of the null hypothesis with the data (categories 3 and 4). The maximum score was therefore 4 for abstracts and 5 for full texts. 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 fundamental statistical aspects (assuming that all other procedures were correct).

The purpose is not to determine the overall methodological rigor but rather the completeness and clarity in applying frequentist-inferential criteria. Indeed, these elements serve as the structural bedrock of the entire model, without which the overall research could be undermined in terms of validity or interpretability. The conditional reliability of the scale was tested by two independent raters on 50 papers. The obtained scores are reported in Table 1.

Table 1. Inter-scale reliability of the SRPS tested on 50 manuscripts by 2 independent raters. The calculations were performed using RStudio, version 4.2.0, 'psych' library, 'cohen.kappa' function.

Category	% agreement	Cohen's k	95% CI	Weighted k	95% CI	null P-value
1	100	1.00	1.00 – 1.00	1.00	1.00 – 1.00	<.0001
2	96	0.91	0.78 – 1.00	0.94	0.86 – 1.00	<.0001
3	94	0.90	0.79 – 1.00	0.94	0.87 – 1.00	<.0001
4	92	0.83	0.68 – 0.98	0.87	0.77 – 0.98	<.0001
5	90	0.79	0.62 – 0.96	0.89	0.79 – 0.99	<.0001

The SRPS was applied to both abstracts and entire papers to highlight possible differences. Indeed, while it is true that abstracts force authors to provide a partial representation of the results, this constraint compels them to give more weight to the information they consider most important. Consequently, this can reveal any interpretative errors and biases. For instance, prior literature indicates that authors often select statistical findings they believe to be most relevant purely based on the P-value for the mathematically null hypothesis, which - as discussed in the previous section - is a practice devoid of scientific value. In addition, reading an abstract could influence the interpretation of the entire manuscript (according to the confirmation bias mechanism) or, in more severe cases, even replace the reading of the manuscript. All data are reported in full at this URL: https://osf.io/3qgcs (DOI: 10.31219/osf.io/3qgcs).

Statistical analysis

The Wilcoxon signed-rank test was employed to compare the scores of the abstracts and those of the full texts. This method was chosen because the data are ordinal and the compared groups are dependent. Random sampling was carried out as described in the subsection "Selection criteria and collection procedure." It was assumed that the sample size was sufficient to adequately represent the reference population (to be confirmed with further studies). At the same time, the validity of the comparison is contingent upon the evaluative ability of the author. For these reasons, the results are described with an unconditional approach. The standardized effect size was measured with the formula $r = z/\sqrt{n}$, where n represents the number of non-zero pairs. The Statistics Kingdom's calculator was adopted in this scope.[21]

Results

Abstracts analysis.

None of the 200 studies adopted the neo-Fisherian approach to describe statistical incompatibility. The vast majority of studies did not report P-values in their entirety (96) or reported them only for results considered statistically significant (85). The vast majority of studies did not report measures of effect size (115) or best estimates (36), or reported them only for results considered statistically significant (71 and 135, respectively). A complete summary is presented in Table 2.

Table 2. This table reports the scores achieved by the 200 abstracts included in the study. Legend: N.A. = not available.

	Compatibility ranges	Full null P-values	Effect size	Best estimates
0 points	200	96	115	36
0.5 points	N.A.	85	71	135
1 point	0	19	14	29

Full-text analysis.

None of the 200 studies adopted the neo-Fisherian approach to describe statistical incompatibility. The vast majority of studies reported P-values in their entirety (141). Regarding effect size and best estimates, the range with the highest score had the most studies (76 and 138, respectively), although a substantial number of these did not report such measures (49 and 1, respectively) or only reported them for values considered statistically significant (75 and 61, respectively). Finally, the vast majority of studies did not mention the basic statistical assumptions of the tests used (163) or merely explained how these were treated without reporting quantitative results or motivated arguments (29). A complete summary is presented in Table 3.

Table 3. This table reports the scores achieved by the 200 full texts included in the study. Legend: N.A. = not available.

	Compatibility ranges	Full null P-values	Effect size	Best estimates	Assumptions
0 points	200	22	49	1	163
0.5 points	N.A.	37	75	61	29
1 point	0	141	76	138	8

Comparison between abstracts and full-texts. The result of the Wilcoxon signed-rank test showed a marked incompatibility of the null hypothesis (of zero difference between the scores of the abstracts and full texts) with the data obtained in the survey (null $P < .001$ as assessed by the test). This outcome warrants further investigation through studies on larger samples and over longer periods. The statistical effect size was considerable ($r = 0.87$). Additionally, the score difference (207.5 points for abstracts versus 441.5 points for full-texts) indicates a scenario consistent with a substantial disparity in the quality and completeness of the outcomes presentation.

Discussion

Principal findings

In light of the associated costs and risks, this paper shows how the misuse of statistical significance is still highly problematic within the field of public health. None of the 200 analyzed studies used the neo-Fisherian approach to properly inform scientific conclusions. At the same time, the dichotomous Wald-Neyman-Pearson approach, ideally employed to limit the total number of incorrect decisions in numerous equivalent experiments (an impossibility to guarantee in complex scientific contexts such as epidemiology or pharmacology), was mistakenly adopted to draw conclusions on individual studies. However, in the vast majority of cases - albeit in the full texts only - all P-values and best estimates were reported in their entirety, allowing the reader to make an independent evaluation. In this regard, the difference in result presentation between the full manuscripts and their abstracts is not solely explainable due to constraints imposed by the latter. On many occasions, even when dealing with a small number of tests, only results with $P < .05$ for the mere null hypothesis were presented. Besides, phrases indicating the statistical non-significance of other findings were reported without showing any P-value or effect size measure. This behavior aligns with the concerning cognitive bias highlighted by the statistical community.[1-6, 15-20, 22] In addition, almost all the studies provided no framework for assessing the validity of the assumptions underlying the tests employed, making the reported outcomes extremely susceptible to compromising margins of interpretative errors. This scenario aligns with the little importance generally given to the background hypotheses of the model employed, which are as essential as the target hypothesis and deserve the same degree of analytical attention.

Practical implications

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 own 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 prolonged funding for studies with exaggerated outcomes.[1, 2, 14, 18, 23, 24]] Sensationalistic expressions to increase the perception of the study's importance beyond its actual findings, i.e., to boost the number of citations and success, are crucial in securing research funding and even institutional roles.[25] For instance, the scientific community has been decrying the widespread practice of P-hacking for decades, although the consequences of this misconduct are a subject of debate.[26, 27] As highlighted by the undersigned and various experts in the field, as well as corroborated by these findings, there is furious resistance to changing these scientifically unsound practices.[28] Therefore, the author of this manuscript calls for academic journals to begin mandating scientific standards that align with the latest statistical evidence advocated by the American Statistical Association.[5] 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. At the same time, albeit with a more limited impact, it is important to stress that misunderstandings regarding statistical testing also afflict major scientific bodies. For instance, in a recent study on COVID-19 vaccines' adverse events, following the recommendations of some experts from the Centers for Disease Control and Prevention and the Global Vaccine Data Network regarding observed versus expected ratios (OE), the authors deemed various results as less relevant only because the lower bound of the 95% compatibility interval was below 1.5.[29] However, cases like Guillain-Barré syndrome led to wide 95% compatibility intervals (0.48, 14.41), with a point estimate consonant with a large effect (OE = 3.99). Conditionally on the model's assumptions, such a scenario is much more compatible with hypotheses of high magnitude than with the mathematically null hypothesis OE = 1 or the small hypothesis OE = 1.5. The width of the above interval conditionally indicates a marked statistical uncertainty and not the absence of significance or relevance. Based on this, the following basic recommendations are proposed. First, if and 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 conditional (in)compatibility between the experimental data and the target hypothesis (e.g., null hypothesis) as assessed by the chosen test. Specifically, P-values close to 1 indicate high compatibility (low incompatibility), while P-values close to 0 indicate low compatibility (high incompatibility). 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 mere statistical analyses. If taken alone, frequentist-inferential statistics is mathematically unable to provide evidence in favor of a real phenomenon since it operates in a utopian world that assumes chance as the sole factor at play.

At most, through the unconditional-descriptive approach, the P-value helps inform conclusions whose foundations must also rely on evidence of other natures (e.g., biochemical, clinical, physical, etc.). Finally, academic journals should explicitly and compulsorily require that public health recommendations be provided only after an analysis of previous literature, data sensitivity, biases, confounding factors, risks, costs, and benefits, and not on the P-value or any other statistical indicator.[16-20, 22, 23, 30, 31] In this regard, guidelines and checklists previously discussed in the literature (e.g., SAMBR) can be helpful.[32-35] Moreover, it's also worth noting that there are international initiatives aimed at improving the quality and transparency of health research, such as EQUATOR.[36]

Proposal

Based on the above, the following guidelines are proposed for reporting basic statistical analyses comprehensively:

1. In the manuscript, provide a brief explanation of how the tests were performed and how the related background assumptions were examined, specifying – if necessary – that methodological details are documented in a supplementary file. Authors' opinions and expectations should be clearly stated in order to openly acknowledge potential biases.
2. Present all calculations and procedures used to validate the adopted tests (including their background assumptions). Quantitative and qualitative results, including graphs, should be fully reported so that readers can independently and easily assess their validity. Indeed, a statistical test is reliable if and only if all the underlying assumptions are true (or sufficiently met). This point can be addressed directly in the manuscript or, if necessary, in a supplementary file.
3. Present, at least in the full text, all P-values and effect size measures, regardless of the supposed "significance" or other properties.
4. Avoid the dichotomous use of the terms “significant” and “non-significant” since they are unscientific. Instead, use the P-value as a continuous measure of (in) compatibility between the experimental data and the target hypotheses as evaluated by the chosen statistical model (after validating its assumptions).
5. Additionally, though not addressed in this study, it is crucial to consider the implementation of more advanced techniques (e.g., adjustment for multiple comparisons and sensitivity analysis) depending on the research purpose and scenario.[31] This includes the adoption of multiple hypotheses (e.g., difference = 0, difference = 0.5, etc.) or multiple compatibility/ confidence intervals (e.g., see the notation 99/95/90%-CI recently proposed).[18, 19] The references mentioned above can be useful for this purpose.

Limitations and their potential impact

This study has some limitations that should be taken into account. Firstly, the sample was collected over a very recent but limited period. There may be trends or periodic oscillations to consider, even though the author is not aware of them and does not find valid reasons to suspect their existence. Secondly, the study focused on the most commonly used statistical approaches in public health but did not consider other methods that might be adopted in this field. However, since i) the goal is to provide a simple overview and ii) the primary statistical models have been included, the author believes that this potential limitation does not have a practical impact. Thirdly, part of the study relied on a newly developed evaluation scale (SRPS), which, while tested for reliability, has not been extensively validated. Nevertheless, the interpretation of this scale is very straightforward and, in any case, allows for easy independent reading. Fourthly, it is possible that some researchers may not have used acronyms (e.g., ANOVA) but instead provided an extended description (e.g., analysis of variance). However, the author is not aware of any behavioral distinctions between those who use acronyms and those who do not. For this reason, he considers the sample to be sufficiently representative in this sense.

Conclusion

These findings align with the hypothesis of widespread and severe shortcomings in the use of statistical significance within public health research during 2023. This scenario is strongly consistent with decades of criticism from known epidemiologists and statisticians, including respected international organizations like the American Statistical Association. Such errors can lead to highly misleading interpretations, thus posing a direct threat to public safety. Therefore, it is essential for academic journals to demand higher scientific quality standards. The suggestions provided in this study could be useful for this purpose.

Conflict of Interests

The author declares that he has no known conflict of interest.

Funding

This study did not receive any funding.

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. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019 Mar;567(7748):305-307. doi: 10.1038/d41586-019-00857-9. PMID: 30894741.
3. Gelman A. The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. Pers Soc Psychol Bull. 2018 Jan;44(1):16-23. doi: 10.1177/0146167217729162. Epub 2017 Sep 7. PMID: 28914154.
4. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. Am Stat. 2019;73(sup1):235–245. doi:10.1080/00031305.2018.1527253.
5. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70(2):129–133. doi:10.1080/00031305.2016.1154108.
6. McShane BB, Bradlow ET, Lynch JG, Meyer RJ. "Statistical Significance" and Statistical Reporting: Moving Beyond Binary. J Marketing. 2024;88(3):1-19. doi:10.1177/00222429231216910.
7. Lehmann EL. Epilog. In: Lehmann EL, editor. Fisher, Neyman, and the creation of classical statistics. New York: Springer New York; 2011. pp. 87–93.
8. Greenland S. Divergence versus decision P-values: A distinction worth making in theory and keeping in practice: Or, how divergence P-values measure evidence even when decision P-values do not. Scand J Statist. 2023;50(1):54–88. doi:10.1111/sjos.12625.
9. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: an explanation for new researchers. Clin Orthop Relat Res. 2010 Mar;468(3):885-92. doi: 10.1007/s11999-009-1164-4. PMID: 19921345; PMCID: PMC2816758.
10. Rovetta A. A Framework to Avoid Significance Fallacy. Cureus. 2023 Jun 11;15(6):e40242. doi: 10.7759/cureus.40242. PMID: 37440801; PMCID: PMC10334213.
11. Rubin M. Repeated sampling from the same population? A critique of Neyman and Pearson's responses to Fisher. Eur Jnl Phil Sci. 2020;10:42. doi:10.1007/s13194-020-00309-6.
12. Greenland S. Connecting simple and precise P-values to complex and ambiguous realities (includes rejoinder to comments on "Divergence vs. decision P-values"). Scand J Statist. 2023;50(3):899–914. doi:10.1111/sjos.12645.
13. Greenland S. Transparency and disclosure, neutrality and balance: shared values or just shared words? J Epidemiol Community Health. 2012 Nov;66(11):967-70. doi: 10.1136/jech-2011-200459. Epub 2012 Jan 19. PMID: 22268131.
14. Greenland S. There's Not Much Science in Science Addressing the Psychosocial Gap in Methodology. April 19th, 2023. Harvard T.H. Chan School of Public Health. [Internet]. Available from: https://www.hsph.harvard.edu/event/theres-not-much-science-in-science-addressing-the-psychosocial-gap-in-methodology/
15. Amrhein V, Trafimow D, Greenland S. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. Am Stat. 2019;73(sup1):262–270. doi:10.1080/00031305.2018.1543137.
16. Amrhein V, Greenland S. Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. J Inf Technol. 2022;37(3):316-320. doi:10.1177/02683962221105904.
17. Greenland S, Mansournia MA, Joffe M. To curb research misreporting, replace significance and confidence by compatibility: A Preventive Medicine Golden Jubilee article. Prev Med. 2022 Nov;164:107127. doi: 10.1016/j.ypmed.2022.107127. Epub 2022 Jul 3. PMID: 35787846.
18. Rovetta A. S-values and Surprisal Intervals to Replace P-values and Confidence Intervals. REVSTAT-Statistical Journal. 2024 Jan 12 [Epub ahead of print]. Available from: https://revstat.ine.pt/index.php/REVSTAT/article/view/669
19. Rovetta A. Common Statistical Errors in Scientific Investigations: A Simple Guide to Avoid Unfounded Decisions. Cureus. 2023 Jan 4;15(1):e33351. doi: 10.7759/cureus.33351. PMID: 36751163; PMCID: PMC9897709.
20. Rovetta A. Multiple Confidence Intervals and Surprisal Intervals to Avoid Significance Fallacy. Cureus. 2024 Jan 9;16(1):e51964. doi: 10.7759/cureus.51964. PMID: 38333481; PMCID: PMC10852995.
21. Statistics Kingdom. Wilcoxon Signed Rank test calculator [Internet]. [cited 2024 Apr 19]. Available from: https://www.statskingdom.com/175wilcoxon_signed_ranks.html
22. Mansournia MA, Nazemipour M, Etminan M. P-value, compatibility, and S-value. Glob Epidemiol. 2022

Sep 12;4:100085. doi: 10.1016/j.gloepi.2022.100085. PMID: 37637018; PMCID: PMC10446114.

23. Amrhein V, Greenland S, McShane BB. Statistical significance gives bias a free pass. Eur J Clin Invest. 2019 Dec;49(12):e13176. doi: 10.1111/eci.13176. Epub 2019 Nov 15. PMID: 31610012.

24. 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.

25. 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.

26. Friese M, Frankenbach J. p-Hacking and publication bias interact to distort meta-analytic effect size estimates. Psychol Methods. 2020 Aug;25(4):456-471. doi: 10.1037/met0000246. Epub 2019 Dec 2. PMID: 31789538.

27. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. 2015 Mar 13;13(3):e1002106. doi: 10.1371/journal.pbio.1002106. PMID: 25768323; PMCID: PMC4359000.

28. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol. 2020 Sep 30;20(1):244. doi: 10.1186/s12874-020-01105-9. PMID: 32998683; PMCID: PMC7528258.

29. Faksova K, Walsh D, Jiang Y, Griffin J, Phillips A, Gentile A, Kwong JC, Macartney K, Naus M, Grange Z, Escolano S, Sepulveda G, Shetty A, Pillsbury A, Sullivan C, Naveed Z, Janjua NZ, Giglio N, Perälä J, Nasreen S, Gidding H, Hovi P, Vo T, Cui F, Deng L, Cullen L, Artama M, Lu H, Clothier HJ, Batty K, Paynter J, Petousis-Harris H, Buttery J, Black S, Hviid A. COVID-19 vaccines and adverse events of special interest: A multinational Global Vaccine Data Network (GVDN) cohort study of 99 million vaccinated individuals. Vaccine. 2024 Apr 2;42(9):2200-2211. doi: 10.1016/j.vaccine.2024.01.100. Epub 2024 Feb 12. PMID: 38350768.

30. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014 Dec;43(6):1969-85. doi: 10.1093/ije/dyu149. Epub 2014 Jul 30. PMID: 25080530.

31. 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.

32. Dwivedi AK. How to write statistical analysis section in medical research. J Investig Med. 2022 Dec;70(8):1759-1770. doi: 10.1136/jim-2022-002479. Epub 2022 Jun 16. PMID: 35710142; PMCID: PMC9726973.

33. Dwivedi AK, Shukla R. Evidence-based statistical analysis and methods in biomedical research (SAMBR) checklists according to design features. Cancer Rep (Hoboken). 2020 Aug;3(4):e1211. doi: 10.1002/cnr2.1211. Epub 2019 Aug 22. PMID: 32794640; PMCID: PMC7941456.

34. Mansournia MA, Collins GS, Nielsen RO, Nazemipour M, Jewell NP, Altman DG, Campbell MJ. A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration. Br J Sports Med. 2021 Sep;55(18):1009-1017. doi: 10.1136/bjsports-2020-103652. Epub 2021 Jan 29. PMID: 33514558; PMCID: PMC9110112.

35. Mansournia MA, Nazemipour M. Recommendations for accurate reporting in medical research statistics. Lancet. 2024 Feb 17;403(10427):611-612. doi: 10.1016/S0140-6736(24)00139-9. PMID: 38368003.

36. EQUATOR Network. Enhancing the QUAlity and Transparency Of health Research. [Internet]. [cited 2024 Apr 19]. Available from: https://www.equator-network.org/