Zeitschrift für
Rheumatologie

Hot Topics

Z Rheumatol 2020 · 79:692–695
https://doi.org/10.1007/s00393-020-00835-x
Online publiziert: 3. Juli 2020

A. Richter[1] · **A. Zink**[2]

[1] Institut für Community Medicine, Universitätsmedizin Greifswald, Greifswald, Deutschland
[2] Deutsches Rheuma-Forschungszentrum Berlin, Ein Institut der Leibniz-Gemeinschaft, Berlin, Deutschland

Gehört die statistische Signifikanz aufs Altenteil?

Hintergrund

Mit der Schlagzeile „Retire statistical significance" haben Amrhein et al. in *Nature* dazu aufgerufen, sich vom Konzept der statistischen Signifikanz in der Medizin zu verabschieden [3]. Unterstützt durch ein Editorial [9] und Unterschriften von mehr als 800 Wissenschaftlern hat diese Arbeit eine rege Diskussion ausgelöst [13, 16, 18–20].

Worum geht es und warum diese Aufregung? Handelt es sich nicht um eine Problematik, die in den vergangenen 50 Jahren immer wieder kritisch diskutiert wurde [10, 11, 15, 17, 23]? Die statistische Signifikanz ist ein Konzept, das uns in der Situation der Unsicherheit Hilfestellung geben soll. Empirische Forschungsergebnisse repräsentieren immer nur einen Ausschnitt aus der Wirklichkeit und sind unter ganz bestimmten Bedingungen entstanden. Die Signifikanztestung hat die einzige Aufgabe, uns Orientierung bei der Bewertung der Ergebnisse zu geben. Verwendet wird die statistische Signifikanz hingegen sehr oft, um Ja/nein-Entscheidungen abzuleiten, selbst wenn bei geringfügig anderen Ergebnissen die gegenteilige Entscheidung gefällt worden wäre. Unter/über einer bestimmten Signifikanzschwelle werden Ergebnisse als bestätigt/abgelehnt angesehen.

Die Kritik am Konzept

Amrhein et al. erinnern an zwei wesentliche Punkte: erstens, dass ein *p*-Wert von >0,05 für ein Studienergebnis oder ein Konfidenzintervall, das die Eins einschließt, nicht bedeutet, dass *kein Unterschied* bzw. *keine Assoziation* bestünde. Sondern nur, dass diese Assoziation nicht gezeigt werden konnte. Zweitens, dass bei identischen Effektmaßen zweier Studien, wie in **Abb. 1** mit einer jeweiligen Hazard Ratio von 1,5 dargestellt, unterschiedliche Konfidenzintervalle der Studien – eines schließt die Eins ein und das andere nicht – keine einander widersprechenden Ergebnisse darstellen [3]. Dies kann aus statistischer und inhaltlicher Perspektive nur unterstützt werden. Die Liste falscher Interpretationen statistisch signifikanter Ergebnisse ließe sich leicht erweitern; Greenland et al. haben hierzu 25 gängige Fehlinterpretationen zusammengetragen [12]. In **Tab. 1** sind weitere mögliche Stolpersteine bei der Interpretation von *p*-Werten erwähnt.

Entgegen dieser häufigen Kritik ist die gelebte wissenschaftliche Praxis anders, und darauf zielen die Autoren ab. *p*-Werte werden u. a. zu Hunderten [4] in Publikationen verwendet und für dichotomisierte Entscheidungen herangezogen. Das Konzept statistischer Signifikanz, zu dem neben *p*-Werten mittelbar auch die Fehler 1. und 2. Art sowie Konfidenzintervalle für Effekte gehören, wird falsch verwendet und falsch verstanden. Die Frage, ob Unterschiede zwischen zwei Gruppen klinisch relevant sind, wird oftmals gar nicht gestellt. Es manifestiert sich hier ein naiver Umgang mit Signifikanz und wissenschaftlicher Unsicherheit.

Den Autoren ist zuzustimmen, dass die starre Fixierung auf *p*-Werte die Unsicherheit außen vor lässt, unter der Studienergebnisse entstehen. Ungenaue, unzureichende Daten oder selektive Studienpopulationen tragen u. a. zur Unsicherheit der Ergebnisse bei. Eine Bewertung dieser muss über *p*-Werte hinaus in einem elaborierten Umgang mit statistischer Unsicherheit erfolgen. Hierzu gehört z. B., ein Ergebnis mit verschiedenen Methoden zu prüfen, Sensitivitätsanalysen durchzuführen und an jeder Stelle nach der klinischen Relevanz zu fragen. Dies gilt ganz besonders für nicht randomisierte Studiendesigns, wo wir mit einer Vielzahl von Verzerrungsfaktoren umgehen müssen. Wasserstein et al. fassen dies unter dem sehr passenden Akronym ATOM zusammen: Akzeptanz der Unsicherheit („*A*ccept uncertainty"), Sorgfalt („be *T*houghtful"), Aufgeschlossenheit („*O*pen") und Zurückhaltung („*M*odest") [28] sollten bei der Interpretation von Studienergebnissen vorliegen.

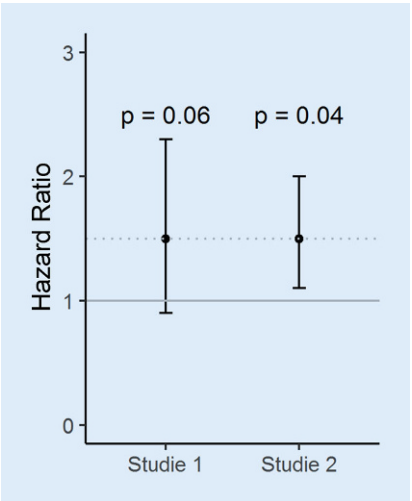

Abb. 1 ▲ Hazard Ratios, Konfidenzintervalle und *p*-Werte aus zwei Studien. (Angelehnt an Amrhein et al. [3])

Tab. 1 Aussagekraft des *p*-Wertes und Erklärungen zur Verwendung des *p*-Wertes in verschiedenen Kontexten

***p*-Wert und:**	**Erklärung**
Aussagekraft	Der *p*-Wert ist ein Ergebnis der Analyse, dessen Größe von vielen Faktoren abhängt, u. a. dem Zutreffen aller Modellannahmen, den vorliegenden Daten und deren Manipulation. Der *p*-Wert ist nur die Wahrscheinlichkeit, unter der Nullhypothese ein noch unwahrscheinlicheres Ergebnis für die Testgröße zu erhalten Beispiel: Ein Studienergebnis weist den Zusammenhang zwischen Alter und systolischem Blutdruck (SBD) mit einem Anstieg um 6,5 mm HG pro Altersdekade (95 %-Konfidenzintervall [6,0; 7,0]) und einem *p*-Wert von $p = 0{,}001$ aus. Unter der Nullhypothese (zwischen Alter und SBD besteht kein Zusammenhang) ist die Wahrscheinlichkeit für ein extremeres Testergebnis 0,1 %
Klinische Relevanz	Ob Studienergebnisse auch eine klinische Relevanz haben, ist eine inhaltliche Entscheidung, die unabhängig von *p*-Werten zu beurteilen ist
Kausalität	Der *p*-Wert allein hat keine Aussagekraft hinsichtlich der Kausalität von Studienergebnissen. Für Aussagen zur Kausalität sind die zugrunde liegende Methodik und das Studiendesign entscheidend. Konfirmatorische, randomisierte Studien lassen Aussagen zur möglichen Kausalität zu
Effektstärke	Von der Größe des *p*-Wertes kann nicht auf die Effektstärke geschlossen werden. Sowohl große wie auch kleine Effekte können mit hohen sowie sehr kleinen *p*-Werten beschrieben werden. Die Größe des *p*-Wertes ist insbesondere von der Fallzahl abhängig. In sehr großen Studien, beispielsweise mit Sekundärdaten, führen selbst marginale Effekte zu sehr kleinen *p*-Werten
Existenz von Effekten	*p*-Werte >0,05 bedeuten nicht, dass ein untersuchter Effekt tatsächlich nicht vorliegt, sondern lediglich, dass er mit den verwendeten Daten nicht gezeigt werden konnte [1]. Falsche Methodik, unzureichende Fallzahl oder ungenaue Daten können dazu führen, dass vorhandene Effekte nicht gezeigt werden können
Multiples Testen	Die häufige Anwendung von mehreren unabhängigen Hypothesentests führt zur Erhöhung der Wahrscheinlichkeit, dass die Nullhypothese abgelehnt wird, obwohl sie korrekt ist. Bei 100 unabhängigen Tests – diese Anzahl wird nicht selten in Publikationen erwähnt [4] – sind 5 Tests falsch positiv (signifikant)
Bias	Die Auswahl von Studienergebnissen aufgrund niedriger *p*-Werte führt zu Publikationsbias, dies ist hinlänglich bekannt Weniger bekannt ist, dass *p*-Werte selbst auch einer Verzerrung (Bias) unterliegen können. In multiplen Regressionsmodellen führt starke Kollinearität zwischen Prädiktoren zur Inflation der Varianz der Schätzer. Damit sind die entsprechenden *p*-Werte überschätzt
Berichterstattung in Publikationen	Die alleinige Darstellung von *p*-Werten in Publikationen ist fast immer unzureichend. Es sollten die Effekte mit zugehörigen Konfidenzintervallen berichtet werden, ggf. ergänzt um *p*-Werte [21]

Ersetzbarkeit des Konzepts

Aber kann der falsche Gebrauch eines Konzepts als Begründung dafür dienen, es komplett über Bord zu werfen? Dies scheint uns im Moment zumindest fraglich. Ein universales und unter Methodikern konsentiertes besseres Konzept existiert nicht [28]. Alternativen wie die 2. Generation von *p*-Werten [6], die Bayes-Faktor-Schranke [5] und einige weitere [28] bedürfen höheren methodischen Verständnisses. Manche dieser Alternativen benötigen weitere Annahmen. Ioannidis schlägt z. B. die Bestimmung des positiv prädiktiven Wertes eines Studienergebnisses vor [17]. Hierfür wird allerdings eine Annahme zum Verhältnis „wahrer" zu „nicht-wahrer" Ergebnisse und zum potenziellen Bias der Studie benötigt [17]. Andere Ansätze [8] sind momentan in gängiger Software zur statistischen Modellierung nicht so implementiert, dass Nicht-Programmierer sie einsetzen könnten. Werden sie eingesetzt, so ist die Gefahr der falschen Anwendung bei diesen komplexeren Verfahren eher als höher einzuschätzen als bei dem eigentlich trivialen Konzept der statistischen Signifikanz mit *p*-Werten und Konfidenzintervallen.

Der Kontext der Anwendung ist entscheidend

Die Verwendung statistischer Tests und das Testen von Hypothesen erfolgen u. a. in verschiedenen Stadien klinischer oder epidemiologischer Studien. Sind die Studien präklinisch, werden sie mit Routinedaten durchgeführt oder liegt ein Beobachtungsdesign zugrunde, so sind sie eher explorativ. Diese Studien sind unverzichtbar für das Generieren von Hypothesen [26], die dann in gut geplanten konfirmatorischen Studien wie randomisierten klinischen Versuchen überprüft werden können. Letztere werden exklusiv zur Untersuchung bestimmter Hypothesen geplant und durchgeführt.

Diese unterschiedlichen Anwendungsbereiche des statistischen Testens diskutieren Amrhein et al. [3] unzureichend und fordern eine globale Abschaffung des Konzepts. Uns erscheint es dagegen durchaus legitim, in einer explorativen Analyse einer Beobachtungsstudie mit vielen möglichen Einflussfaktoren auch einen statistisch nicht signifikanten Zusammenhang zu beschreiben. Ebenso ist bei der statistischen Modellierung die Wahl der am besten zu den Daten passenden Verteilungsform, z. B. negativ-binomial vs. Poisson-Verteilung, basierend auf dem *p*-Wert eines Likelihood-Ratio-Tests, hinreichend. Für den Fall der klinischen Prüfstudie allerdings, die für einen Effektivitätsparameter als primären Endpunkt geplant wurde, ist es keine gute wissenschaftliche Praxis, die Risikoerhöhung eines Therapiearms um 50 % (**Abb. 1**, Studie 1) auf fehlende statistische Signifikanz zu reduzieren, nur weil der *p*-Wert >0,05 ist. Die Evidenz für das Nichtvorliegen einer Risikoerhöhung kann nicht allein vom *p*-Wert abgeleitet werden. Altman fasste dies in den Worten zusammen: „die Abwesenheit von Evidenz" (für einen Effekt, durch *p*-Werte abgeleitet) „ist keine Evidenz für Abwesenheit" eines Effekts [1]. Korrekt wäre in der oben erwähnten Studie die Schlussfolgerung, dass sie für diesen Endpunkt nicht geplant wurde und diesbezüglich keine ausreichende statistische Power aufweist und dass die beobachtete Risikoerhöhung, insofern sie sich auf

einen klinisch bedeutsamen Endpunkt bezieht, einer weiteren Überprüfung in adäquat geplanten Studien bedarf.

Mehr Verzerrung und weniger Transparenz

Letzteres Beispiel veranschaulicht zugleich einen wesentlichen Vorteil des Konzepts statistischer Signifikanz: Eine inkorrekte Interpretation ist für den Leser oder Zuhörer zumeist transparent. Wenn Effektstärken, Konfidenzintervalle und *p*-Werte gemeinsam dargestellt werden, können die Schlussfolgerungen hinterfragt werden. Dieses Merkmal des Konzepts der statistischen Signifikanz ist vor dem Hintergrund zunehmender Ergebnisse aus Messverfahren wie „next-generation-sequencing" oder Machine-learning-Modellen hoch zu bewerten, da ihre Generierung und Fehleranfälligkeit für Leser weitgehend intransparent sind [7, 22, 25].

Die komplette Abschaffung der Signifikanztestung würde der willkürlichen Interpretation von Studienergebnissen Tür und Tor öffnen [16]. Vor allem, weil ein gutes Studiendesign verlangt, dass die Kriterien für einen relevanten Unterschied vorab und nicht nach Datenlage festgelegt werden. Dies führt uns zurück zu Grundprinzipien wissenschaftlicher Arbeit wie in der „Guten Epidemiologischen Praxis" beschrieben [14]. Die Diskussion kann daher nicht sein, *ob*, sondern *wie* getestet wird und wie die Ergebnisse interpretiert werden.

In diesem Zusammenhang muss die Forderung nach der vollständigen Publikation auch nichtsignifikanter Ergebnisse unterstrichen werden. Wenn Metaanalysen nicht durch den Publikationsbias (also die bevorzugte Publikation signifikanter Ergebnisse) verzerrt sind, können fälschliche Schlussfolgerungen einzelner Studien in einem globaleren Kontext bewertet und in einer Gesamtbewertung evtl. korrigiert werden. Dies gilt nicht nur für Ergebnisse klinischer Studien, sondern auch für präklinische Studien, wo oftmals sehr viele verschiedene Biomarker geprüft werden und ohne Korrektur für multiples Testen leicht falsch positive oder falsch negative Ergebnisse entstehen [24].

Zusammenfassung · Abstract

Z Rheumatol 2020 · 79:692–695 https://doi.org/10.1007/s00393-020-00835-x

A. Richter · A. Zink

Gehört die statistische Signifikanz aufs Altenteil?

Zusammenfassung

Unter der Schlagzeile „Retire statistical significance" haben Amrhein et al. in der Zeitschrift *Nature* dazu aufgerufen, sich vom Konzept der statistischen Signifikanz zu verabschieden. Dieser von rund 800 weiteren Forschern unterzeichnete Aufruf löste eine kontroverse Diskussion aus. Ein Grund für die bewusst provokante Forderung ist die gelebte wissenschaftliche Praxis, in der das Konzept der statistischen Signifikanz häufig eine falsche Anwendung findet, indem sie für Ja/nein-Entscheidungen herangezogen wird. Die Kritik ist nicht neu und wurde in den letzten 50 Jahren wiederholt geäußert. Wir verweisen auf aktuelle und zurückliegend publizierte Vorbehalte, geben einen Überblick über unterschiedliche Anwendungen des Konzepts der statistischen Signifikanz sowie mögliche Alternativen. Der durch Amrhein et al. geäußerten Kritik am Konzept ist grundsätzlich zuzustimmen. Mangels konsentierter Alternativen und einer zu geringen Berücksichtigung der vielen verschiedenen Anwendungsfälle des Konzepts der statistischen Signifikanz sehen wir die Forderung nach ihrer Abschaffung aber als überzogen an. Ein pragmatischerer Umgang mit der Problematik, unterstützt durch gezielte Handreichungen für Wissenschaftler und Reviewer, erscheint uns der geeignetere Weg.

Schlüsselwörter

P-Wert · Konfidenzintervalle · Hypothesen · Statistische Tests · Klinische Relevanz

Should statistical significance be retired?

Abstract

In the journal *Nature*, under the headline "Retire statistical significance", Amrhein et al. called for the concept of statistical significance to be abolished. This appeal, which was signed by about 800 other researchers, triggered a controversial discussion. One reason for the deliberately provocative call is the scientific practice in which the concept of statistical significance is often applied in an incorrect way for yes/no decisions. The criticism is not new and has been repeatedly expressed over the last 50 years. We refer to current and previously published caveats, give an overview of different applications of the concept of statistical significance as well as possible alternatives. We agree in principle with the criticism of the concept expressed by Amrhein et al. but in the absence of agreed alternatives and insufficient consideration of the many different applications of the concept of statistical significance, we consider the demand for its abolition to be exaggerated. A more pragmatic approach to the problem, supported by targeted instructions for scientists and reviewers, seems to be a more appropriate way forward.

Keywords

p-value · Confidence intervals · Hypotheses · Statistical tests · Clinical relevance

Fazit

Die Forderung nach der Abschaffung des Konzepts statistischer Signifikanz ist derzeit mangels Alternativen und aufgrund unzureichender Differenzierung des Anwendungskontextes überzogen. *p*-Werte werden nicht nur zur Prüfung von Gruppenunterschieden in klinischen Prüfungen herangezogen, sondern auch in explorativen Studien, bei der Wahl eines statistischen Modells oder in Sensitivitätsanalysen zur Bewertung verschiedener Modellannahmen. Wir gehen davon aus, dass dies auch den Autoren bewusst ist. Sie schreiben selbst, dass sie kein Verbot von *p*-Werten und Konfidenzintervallen fordern. Worauf sie eigentlich hinweisen, ist der falsche Gebrauch und die kritiklose Handhabung der statistischen Signifikanz in vielen Studien, und hier ist ihnen uneingeschränkt zuzustimmen.

Eine Handreichung zur Interpretation von Ergebnissen wissenschaftlicher Publikationen, ähnlich der von Lyderson [21], wäre ein konstruktiverer Umgang mit der Problematik als die Abschaffung der Signifikanztestung. Denk-

bar wäre es, Leitlinien wie STROBE [27] oder CONSORT [2] um Checklisten zum Umgang mit statistischem Testen zu ergänzen. Greenland et al. haben nicht nur eine Liste möglicher Fehlinterpretationen, sondern auch Hinweise zur richtigen Interpretation vorgelegt [12]. Auch die Kollegen um Wasserstein et al. fordern mehr Hilfestellung zur richtigen Anwendung und weniger Verbote („‚Don't' is not enough") [28]. Auch wir erachten dies nicht nur für die Autoren von wissenschaftlichen Publikationen, sondern auch für deren Reviewer als hilfreich.

Korrespondenzadresse

Dr. A. Richter
Institut für Community Medicine, Universitätsmedizin Greifswald
Walther-Rathenau-Str. 48, 17475 Greifswald, Deutschland
adrian.richter@uni-greifswald.de

Funding. Open Access funding provided by Projekt DEAL.

Interessenkonflikt. A. Richter und A. Zink geben an, dass kein Interessenkonflikt besteht.

Literatur

1. Altman DG, Bland JM (1995) Statistics notes: absence of evidence is not evidence of absence. BMJ 311:485
2. Altman DG, Schulz KF, Moher D et al (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134:663–694
3. Amrhein V, Greenland S, Mcshane B (2019) Scientists rise up against statistical significance. Nature 567:305–307
4. Anderson DR, Burnham KP, Gould WR et al (2001) Concerns about finding effects that are actually spurious. Wildl Soc Bull 29(1):311–316
5. Benjamin DJ, Berger JO (2019) Three recommendations for improving the use of p-values. Am Stat 73:186–191
6. Blume JD, Mcgowan LDA, Dupont WD et al (2018) Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. PLoS ONE 13:e188299. https://doi.org/10.1371/journal.pone.0188299
7. Chakraborty S, Tomsett R, Raghavendra R et al (2017) Interpretability of deep learning models: a survey of results. In: 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI), S 1–6
8. Colquhoun D (2019) The false positive risk: a proposal concerning what to do about p-values. Am Stat 73:192–201
9. Editorial (2019) It's time to talk about ditching statistical significance. Nature 567:283
10. Feinstein AR (1998) P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 51:355–360
11. Gardner MJ, Altman DG (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 292:746–750
12. Greenland S, Senn SJ, Rothman KJ et al (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31:337–350
13. Haaf JM, Ly A, Wagenmakers E-J (2019) Retire significance, but still test hypotheses. Nature 567:461. https://doi.org/10.1038/d41586-019-00972-7
14. Hoffmann W, Latza U, Baumeister SE et al (2019) Guidelines and recommendations for ensuring Good Epidemiological Practice (GEP): a guideline developed by the German Society for Epidemiology. Eur J Epidemiol 34:301–317
15. Hubbard R, Lindsay RM (2008) Why P values are not a useful measure of evidence in statistical significance testing. Theory Psychol 18:69–88
16. Ioannidis JP (2019) Retiring statistical significance would give bias a free pass. Nature 567:461–462
17. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2:e124
18. Johnson VE (2019) Raise the bar rather than retire significance. Nature 567:461
19. Kramer M (2019) Stats: is this therapy useful? Nature 569:192–192
20. Lepore FE (2019) Stats: 800 signatories on death warrant is overkill. Nature 569:487
21. Lydersen S (2014) Statistical review: frequently given comments. Ann Rheum Dis. https://doi.org/10.1136/annrheumdis-2014-206186
22. Nielsen R, Paul JS, Albrechtsen A et al (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451
23. Petersen R (1977) Use and misuse of multiple comparison procedures 1. Agron J 69:205–208
24. Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10:712
25. Robasky K, Lewis NE, Church GM (2014) The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 15:56–62
26. Tukey JW (1980) We need both exploratory and confirmatory. Am Stat 34:23–25
27. Von Elm E, Altman DG, Egger M et al (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 147:573–577
28. Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a World Beyond "p< 0.05". The American Statistician 73(sup1):1–19