ÜBERSICHTSARBEIT

Konfidenzintervall oder p-Wert?

Teil 4 der Serie zur Bewertung wissenschaftlicher Publikationen

Jean-Baptist du Prel, Gerhard Hommel, Bernd Röhrig, Maria Blettner

ZUSAMMENFASSUNG

Einleitung: Kenntnisse zu p-Werten und Konfidenzintervallen sind zur Beurteilung wissenschaftlicher Artikel notwendig. Dieser Artikel will den Leser über die Bedeutung und Interpretation beider statistischen Konzepte informieren.

Methode: Auf der Grundlage einer selektiven Literaturrecherche zur Methodik in wissenschaftlichen Artikeln wird der Stellenwert von und die Unterschiede zwischen beiden statistischen Konzepten in einer Übersicht dargelegt.

Ergebnisse/Diskussion: Der p-Wert ermöglicht in Studien eine Entscheidung zur Verwerfung oder Beibehaltung einer vorab formulierten Nullhypothese. In explorativen Studien lässt er statistisch auffällige Ergebnisse erkennen. Konfidenzintervalle liefern Informationen über einen Bereich, in dem der wahre Wert mit einer gewissen Wahrscheinlichkeit liegt sowie über Effektrichtung und -stärke. Damit werden Aussagen zur statistischen Plausibilität und klinischen Relevanz der Studienergebnisse möglich. Die Angabe beider statistischen Maße in wissenschaftlichen Artikeln ist oft sinnvoll, da sie einander ergänzende Informationen enthalten.

Dtsch Arztebl 2009; 106(19): 335–9
DOI: 10.3238/arztebl.2009.0335

Schlüsselwörter: Publikation, klinische Forschung, p-Wert, Statistik, Konfidenzintervall

Johannes Gutenberg-Universität Mainz: Zentrum für Kinder- und Jugendmedizin, Zentrum Präventive Pädiatrie: Dr. med. du Prel, MPH

Johannes Gutenberg-Universität Mainz: Institut für Medizinische Biometrie, Epidemiologie und Informatik: Prof. Dr. rer. nat. Hommel, Dr. rer. nat. Röhrig, Prof. Dr. rer. nat. Blettner

Leser wissenschaftlicher Artikel müssen sich bei der Beurteilung der Ergebnisse statistischer Auswertungen mit der Interpretation von p-Werten und Konfidenzintervallen (Vertrauensbereichen) befassen. Mancher wird sich schon gefragt haben, warum in einigen Untersuchungen als Maß der statistischen Wahrscheinlichkeit ein p-Wert angegeben wird, in anderen aber ein Vertrauensbereich, mitunter auch beide. Auf der Grundlage einer selektiven Literaturrecherche erklären die Autoren die beiden Maße und beschreiben, wann p-Werte oder Konfidenzintervalle angegeben werden sollen. Es folgen ein Vergleich und die Beurteilung beider statistischen Konzepte.

Was ist ein p-Wert?

In konfirmatorischen (Beweis führenden) Studien werden Nullhypothesen formuliert, die mithilfe von statistischen Tests verworfen oder beibehalten werden. Beim p-Wert handelt es sich um eine Wahrscheinlichkeit, die das Ergebnis eines solchen statistischen Tests ist. Diese Wahrscheinlichkeit gibt das Ausmaß der Evidenz gegen die Nullhypothese wieder. Kleine p-Werte stellen eine starke Evidenz dar. Ab einem bestimmten p-Wert werden die Ergebnisse als „statistisch signifikant“ bezeichnet (1). In explorativen Untersuchungen spricht man auch von „statistisch auffälligen Ergebnissen“.

Soll gezeigt werden, dass ein neues Medikament besser als ein altes ist, so gilt es zunächst zu beweisen, dass beide Medikamente nicht gleich sind. Die Hypothese der Gleichheit soll also abgelehnt werden. Daher wird die Nullhypothese (H_0), die abgelehnt werden soll, in diesem Fall wie folgt formuliert: „Es gibt keinen Unterschied (Effekt) zwischen den beiden Behandlungen“, zum Beispiel zeigen zwei Antihypertonika keinen Unterschied in ihrer blutdrucksenkenden Wirkung. Die Alternativhypothese (H_1) besagt dann, dass es einen Unterschied zwischen den beiden Therapien gibt. Dabei kann die Alternativhypothese zweiseitig (Unterschied) oder aber einseitig (positiver oder auch negativer Effekt) formuliert werden. Einseitig heißt in diesem Fall, dass man bei Formulierung der Alternativhypothese Vorgaben bezüglich der Richtung des erwarteten Effekts macht. Hat man etwa aus Voruntersuchungen schon deutliche Hinweise dafür, dass ein Antihypertonikum im Mittel eine stärker blutdrucksenkende Wirkung hat als das zu Vergleichende, kann man die Alternativhypothese beispielsweise so formulieren: „Die Differenz der mittleren

Blutdrucksenkung von Antihypertonikum 1 und der mittleren Blutdrucksenkung von Antihypertonikum 2 ist positiv". Da hierzu aber plausible Annahmen hinsichtlich der Effektrichtung erforderlich sind, wird die Hypothese oft zweiseitig formuliert.

Beispielsweise soll aus Daten einer randomisierten klinischen Studie das für die Fragestellung relevante Effektmaß, zum Beispiel die Differenz der mittleren Blutdrucksenkung zwischen einem neuen und dem etablierten Antihypertonikum geschätzt werden. Darauf basierend wird die vorab formulierte Nullhypothese mithilfe eines Signifikanztests überprüft. Der p-Wert gibt dann die Wahrscheinlichkeit an, mit der man das vorliegende Testergebnis oder ein noch extremeres erhält, wenn die Nullhypothese richtig ist. Ein kleiner p-Wert besagt, dass die Wahrscheinlichkeit, dass der Unterschied alleine dem Zufall zugeschrieben werden kann, klein ist. Eine beobachtete Differenz des mittleren systolischen Blutdrucks in unserem Beispiel könnte nicht auf einem echten Unterschied in der blutdrucksenkenden Wirkung der beiden Antihypertonika beruhen, sondern zufällig sein. Bei einem p-Wert $< 0{,}05$ liegt die Wahrscheinlichkeit dafür allerdings unter 5 %. Um eine Entscheidung zwischen Nullhypothese und Alternativhypothese zu ermöglichen, wird vorab oft eine sogenannte Signifikanzgrenze auf einem Signifikanzniveau α festgelegt. Häufig wird ein Signifikanzniveau von 0,05 (beziehungsweise 5 %) gewählt. Unterschreitet der p-Wert diesen Grenzwert (= signifikantes Ergebnis), wird vereinbarungsgemäß die Nullhypothese verworfen und die Alternativhypothese („es gibt einen Unterschied") angenommen. Mit Festlegung des Signifikanzniveaus ist auch die Wahrscheinlichkeit vorgegeben, die Nullhypothese zu Unrecht abzulehnen.

p-Werte alleine erlauben keine direkte Aussage über die Richtung oder Größe einer Differenz oder eines relativen Risikos zwischen unterschiedlichen Gruppen (1). Das wäre aber insbesondere dann nützlich, wenn Ergebnisse nicht signifikant sind (2). Hier beinhalten Vertrauensbereiche mehr Informationen. Neben p-Werten muss zumindest ein Maß der Effektstärke (zum Beispiel Differenz der mittleren Blutdrucksenkung in zwei Behandlungsgruppen) berichtet werden (3). Die Definition einer Signifikanzgrenze ist letztendlich willkürlich und die Angabe von p-Werten ist auch ohne Wahl dieser Signifikanzgröße sinnvoll. Je kleiner der p-Wert ist, umso weniger plausibel wird die Nullhypothese, dass es keinen Unterschied zwischen den Behandlungsgruppen gibt.

Vertrauensbereich – Von der dichotomen Test-Entscheidung zum Effektbereichsschätzer

Ein Vertrauensbereich (Konfidenzintervall) ist ein mithilfe statistischer Methoden berechneter Wertebereich, der den gesuchten, wahren Parameter (zum Beispiel arithmetisches Mittel, Differenz zweier Mittelwerte, Odds Ratio) mit einer vorab definierten Wahrscheinlichkeit (Überdeckungswahrscheinlichkeit, Vertrauenswahrscheinlichkeit oder Konfidenzniveau) überdeckt. Meist wird ein Konfidenzniveau von 95 % gewählt. Das bedeutet, dass in 95 von 100 durchgeführten Studien das Konfidenzintervall den wahren Wert überdecken wird (4, 5). Vorteil der Konfidenzintervalle im Vergleich zu p-Werten ist, dass Konfidenzintervalle die Ergebnisse auf der Ebene der Datenmessung wiedergeben (6). In unserem Beispiel werden etwa die untere und obere Konfidenzgrenze der mittleren systolischen Blutdruckdifferenz zwischen beiden Therapiegruppen ebenfalls in mmHg angegeben.

Die Weite des Vertrauensbereichs hängt von Stichprobengröße und Standardabweichung der untersuchten Gruppen ab (5). Eine große Stichprobe führt zu „mehr Vertrauen" also zu einem engen Konfidenzintervall. Ein breites Konfidenzintervall kann von einer kleinen Stichprobe herrühren. Bei großer Streuung der Werte wird die Aussage unsicherer, das heißt, das Konfidenzintervall wird breiter. Schließlich trägt die Wahl des Konfidenzniveaus zur Weite des Konfidenzintervalls bei. Ein 99-%-Vertrauensbereich ist breiter als ein 95-%-Vertrauensbereich. Oder allgemeiner formuliert: Je mehr Sicherheit man garantieren möchte, desto weiter wird der Vertrauensbereich.

Konfidenzintervalle geben im Unterschied zum p-Wert Aufschluss über die Richtung des zu untersuchenden Effekts. Rückschlüsse auf die statistische Signifikanz sind mithilfe des Konfidenzintervalls möglich. Enthält ein Vertrauensbereich den Wert des „Null-Effekts" nicht, so kann man von einem „statistisch signifikanten" Ergebnis ausgehen. Im Beispiel mit der Differenz des mittleren systolischen Blutdrucks zwischen beiden Therapiegruppen ist die Frage, ob der Wert „0 mmHg" innerhalb (= nicht signifikant) oder außerhalb (= signifikant) des 95-%-Konfidenzintervalls liegt. Entsprechend gilt für das RR (relatives Risiko), dass ein KI, das die 1 enthält, einem nicht signifikanten Ergebnis entspricht. Zu unterscheiden wäre dann, ob das Konfidenzintervall für das relative Risiko vollständig unterhalb der 1 liegt (= protektiver Effekt) oder vollständig oberhalb (= Risikoerhöhung).

Grafik 1 zeigt den Zusammenhang am Beispiel der mittleren systolischen Blutdruckdifferenz zwischen zwei Kollektiven. Das Konfidenzintervall der mittleren Blutdruckdifferenz wird schmal bei kleiner Variabilität innerhalb der Stichproben (= kleine Streuung) *(Grafik 1b)*, kleiner Vertrauenswahrscheinlichkeit *(Grafik 1d)* und großer Fallzahl *(Grafik 1f)*. In diesem Beispiel unterscheiden sich bei großer Streuung *(Grafik 1c)*, hohem Konfidenzniveau *(Grafik 1e)* oder kleiner Fallzahl *(Grafik 1g)* die mittleren systolischen Blutdrucke nicht mehr signifikant, da der Wert Null im Konfidenzintervall enthalten ist.

Punktschätzer (zum Beispiel arithmetisches Mittel, Differenz zweier Mittelwerte oder Odds ratio) liefern zwar die beste Annäherung an den wahren Wert, jedoch keine Information darüber, wie genau sie sind. Dazu dienen Vertrauensbereiche. Exakte Angaben darüber, wie stark der geschätzte Parameter der Stichprobe vom wahren Wert der Grundgesamtheit abweicht, sind natürlich nicht möglich, weil der wahre Wert unbekannt

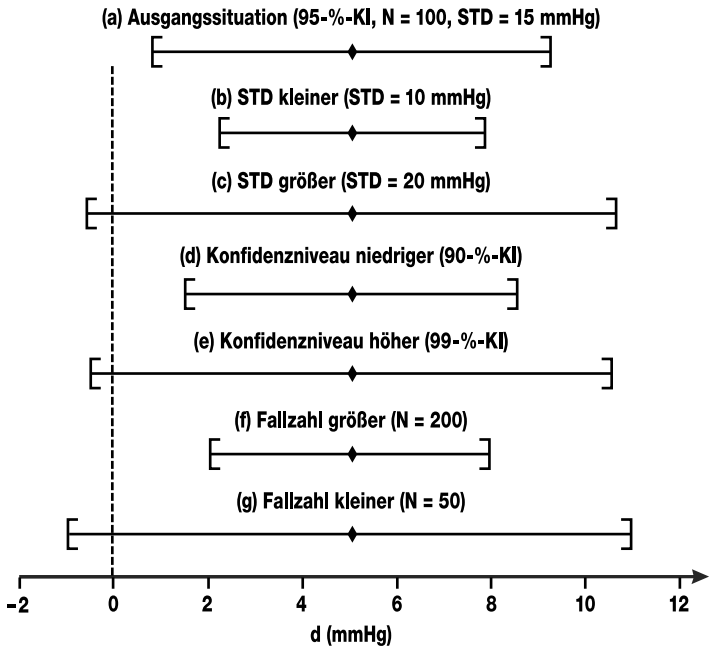
ist. Man möchte aber gerne eine gewisse Sicherheit darüber haben, dass sich der Schätzwert in der Nähe des wahren Wertes befindet (7). Konfidenzintervalle eignen sich zur Beschreibung der Wahrscheinlichkeit, in welchem Bereich sich der wahre Wert befindet.

Durch Angabe eines Vertrauensbereichs lassen sich mehrere Schlüsse ableiten: Zunächst sind Werte unterhalb der unteren beziehungsweise oberhalb der oberen Konfidenzgrenze nicht ausgeschlossen, aber unwahrscheinlich. Bei Verwendung eines 95-%-Vertrauensbereichs beträgt die Wahrscheinlichkeit jeweils nur 2,5 %. Werte, die innerhalb des Vertrauensbereichs, aber nahe der Vertrauensgrenzen liegen, sind meist weniger wahrscheinlich als Werte, die nahe dem Punktschätzer (in unserem Beispiel mit den beiden Antihypertonika wäre das der Mittelwertsunterschied der Blutdrucksenkung in beiden Behandlungsgruppen in mmHg) liegen. Unabhängig von der Weite des Konfidenzintervalls, ist der Punktschätzer auf der Grundlage der Stichprobe die beste Annäherung an den wahren Wert der Grundgesamtheit. Werte in der Nähe des Punktschätzers sind meist plausible Werte. Das gilt insbesondere dann, wenn man eine Normalverteilung der Werte zugrunde legen kann.

Es ist zwar häufige Praxis, Vertrauensbereiche nach dem Kriterium, ob sie eine bestimmte Grenze einschließen oder nicht, ausschließlich hinsichtlich eines signifikanten Ergebnisses zu beurteilen. Besser ist es aber, die genannten zusätzlichen Informationen von Konfidenzintervallen zu nutzen und gerade bei „knappen" Ergebnissen, die Möglichkeit eines signifikanten Ergebnisses bei höherer Fallzahl in die Beurteilung der Ergebnisse mit einzubeziehen.

Bedeutende internationale medizinisch-wissenschaftliche Journals wie „Lancet" oder „British Medical Journal" wie auch das Internationale Komitee der Journaleditoren empfehlen die Verwendung von Vertrauensbereichen (6). Insbesondere bei der Beurteilung von randomisierten, klinischen Studien und Metaanalysen helfen Konfidenzintervalle wesentlich bei der Interpretation der Ergebnisse. So wird in internationalen Vereinbarungen wie dem CONSORT-Statement (8) für die Berichterstattung in randomisierten, klinischen Studien und dem QUORUM-Statement (9) für die Berichterstattung in systematischen Reviews und Metaanalysen die Verwendung von Konfidenzintervallen ausdrücklich gefordert.

GRAFIK 1



KI, Konfidenzintervall; N, Stichprobenumfang pro Gruppe; STD, Standardabweichung je Gruppe als Maß der Streuung (in beiden Gruppen gleich); d, Differenz des systolischen Blutdrucks zwischen beiden Gruppen

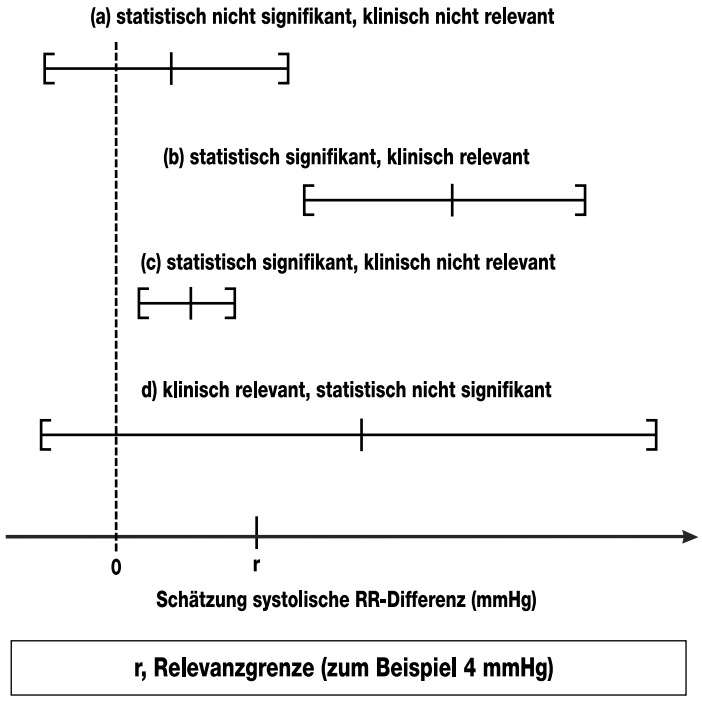
Am Beispiel der Differenz des mittleren systolischen Blutdrucks zwischen zwei Gruppen wird untersucht, wie sich die Breite des Konfidenzintervalls (a) bei Modifizierung von Streuung (b, c), Konfidenzniveau (d, e) und Stichprobenumfang (f, g) verändert. Die Differenz des mittleren systolischen Blutdruckes von Gruppe 1 (150 mmHg) und Gruppe 2 (145 mmHg) lag bei 5 mmHg; Beispiel modifiziert nach (6)

Statistische Signifikanz versus klinische Relevanz

Zwischen statistischer Signifikanz („statistical significance") und klinischer Relevanz („clinical significance") muss man klar unterscheiden. Neben der Effektstärke gehen in p-Werte auch die Fallzahl und die Variabilität der Daten in der Stichprobe ein. Ein vorab festgelegter Grenzwert der statistischen Signifikanz erspart es dem Leser nicht, statistisch signifikante Ergebnisse hinsichtlich ihrer klinischen Relevanz zu beurteilen. Der gleiche numerische Wert für die Differenz kann „statistisch signifikant" bei Wahl einer großen Stichprobe und „nicht signifikant" bei kleiner Stichprobe sein. Andererseits sind Ergebnisse mit hoher klinischer Relevanz aufgrund fehlender statistischer Signifikanz nicht automatisch bedeutungslos. Ursächlich könnte hier eine zu kleine Stichprobe oder eine zu große Streuung der Daten (zum Beispiel durch eine sehr heterogene Patientengruppe) sein. Deshalb ist die Entscheidung auf Basis des p-Wertes in signifikant oder nicht signifikant oft zu einfach.

Das sei am Beispiel mit der systolischen Blutdruckdifferenz verdeutlicht: In *Grafik 2* wird eine Relevanzgrenze r festgelegt: Ein systolischer Blutdruckunterschied von mindestens 4 mmHg zwischen den beiden Behandlungsgruppen wird damit als klinisch relevant definiert. Wenn der Blutdruckunterschied dann weder statistisch signifikant noch klinisch relevant *(Grafik 2a)* oder aber statistisch signifikant und klinisch relevant *(Grafik 2b)* ist, fällt die Interpretation leicht. Statistisch signifikante Blutdruckunterschiede können aber auch unter der klinischen Relevanzgrenze liegen und sind

GRAFIK 2



Statistische Signifikanz und klinische Relevanz

dann klinisch bedeutungslos *(Grafik 2c)*. Andererseits können echte Unterschiede im systolischen Blutdruck zwischen den Behandlungsgruppen mit hoher klinischer Relevanz trotz fehlender statistischer Signifikanz *(Grafik 2d)* gegebenenfalls bedeutungsvoll sein.

Leider wird oft statistische Signifikanz mit klinischer Relevanz gleichgesetzt. Viele Forscher, Leser und auch Journals schenken klinisch potenziell nützlichen Ergebnissen nur deswegen keine Aufmerksamkeit, weil sie statistisch nicht signifikant sind (4). An dieser Stelle sei die Praxis einiger wissenschaftlicher Journals kritisiert, signifikante Ergebnisse bevorzugt zu veröffentlichen. Nach einer Untersuchung war das vor allem bei Journals mit hohem Impactfaktor zu beobachten (10). Dies führt zu einer einseitigen Verzerrung tatsächlicher Begebenheiten („Publikationsbias"). Häufig ist zudem zu beobachten, dass ein nicht signifikantes Ergebnis in klinischen Studien so interpretiert wird, dass es keinen Unterschied, zum Beispiel zwischen zwei Therapiegruppen, gibt. Ein p-Wert von > 0,05 besagt lediglich, dass die Evidenz nicht ausreicht, die Nullhypothese (zum Beispiel unterscheiden sich zwei alternative Therapien nicht) zu verwerfen. Das bedeutet aber nicht, dass beide Therapien deswegen äquivalent sind. Die quantitative Zusammenfassung von vergleichbaren Studien in Form von systematischen Reviews oder Metaanalysen kann dann weiterhelfen, aufgrund einer zu niedrigen Fallzahl nicht erkannte Unterschiede aufzudecken. Diesem Thema ist ein eigener Artikel in dieser Serie gewidmet.

p-Wert versus Konfidenzintervall – Was sind die Unterschiede?

Die wesentlichen Unterschiede zwischen p-Werten und Vertrauensbereichen sind:

- Der Vorteil von Konfidenzintervallen im Vergleich zur Angabe von p-Werten nach Hypothesentestung ist, dass Ergebnisse direkt auf der Ebene der Datenmessung angegeben werden. Konfidenzintervalle geben Informationen sowohl über die statistische Signifikanz als auch über die Richtung und Größe des Effekts (11). Damit kann man auch über die klinische Relevanz der Ergebnisse entscheiden. In die Breite des Vertrauensbereichs bei vorgegebener Irrtumswahrscheinlichkeit gehen zudem die Variabilität der Daten und die Fallzahl der untersuchten Stichprobe ein (12).
- p-Werte sind übersichtlicher als Konfidenzintervalle. Ein Wert kann hinsichtlich des Über- oder Unterschreitens eines vorher bestimmten Grenzwertes beurteilt werden. Damit wird eine schnelle Entscheidungsfindung in statistisch signifikant oder nicht signifikant möglich. Eine solche „Blickdiagnose" kann aber auch dazu verleiten, eine klinische Entscheidung nur unter statistischen Gesichtspunkten zu treffen.
- Die Reduktion der statistischen Inferenz (= induktives Schließen von einer Stichprobe auf die Grundgesamtheit) auf einen Prozess der binären Entscheidungsfindung, wie das bei Hypothesentestung mithilfe des p-Wertes geschieht, kann zu einfach sein. Die reine Unterscheidung zwischen „signifikant" oder „nicht signifikant" ist für sich genommen noch nicht sehr aussagekräftig. Bezüglich der Evidenzlage unterscheidet sich zum Beispiel ein p-Wert von 0,04 nicht viel von einem p-Wert von 0,06. Durch eine binäre Entscheidungsfindung werden aufgrund solcher geringen Unterschiede aber gegenläufige Schlüsse gezogen (1, 13). Aus diesem Grund sollten p-Werte immer vollständig (Vorschlag: immer mit drei Dezimalstellen) angegeben werden (14).
- Mit Punktschätzern (zum Beispiel Mittelwertsdifferenz, relatives Risiko) wird mit nur einem einzigen Wert versucht von der Stichprobe auf die Situation in der Zielpopulation zu schließen. Wenn diese Zahl auch die bestmögliche Annäherung an den wahren Wert ist, so ist eine exakte Übereinstimmung nicht sehr wahrscheinlich. Konfidenzintervalle liefern hingegen einen Bereich mit möglichen plausiblen Werten für die Zielpopulation und eine Wahrscheinlichkeit mit der dieser Bereich den wahren Wert überdeckt.
- p-Werte geben im Unterschied zu Konfidenzintervallen den Abstand von einem vorher festgelegten statistischen Grenzwert, dem Signifikanzniveau α, an (15). Damit fällt die Beurteilung eines „knappen" Ergebnisses leicht.
- Statistische Signifikanz ist von medizinischer Relevanz oder biologischer Bedeutsamkeit zu unterscheiden: Durch Wahl einer genügend großen

Stichprobe können auch sehr kleine Unterschiede statistisch signifikant sein (16, 17). Andererseits können auch große Unterschiede bei unzureichender Fallzahl zu nicht signifikanten Ergebnissen führen (12). In klinischen Studien sollten die Untersucher aufgrund der Bedeutung für den späteren Behandlungserfolg aber mehr an der Größe eines Unterschieds im Therapieeffekt zwischen zwei Behandlungsgruppen interessiert sein, als nur daran ob ein signifikantes oder nicht signifikantes Ergebnis vorliegt (18).

Schlussfolgerung

p-Werte alleine liefern ein Maß für die statistische Plausibilität eines Unterschieds. In Verbindung mit einem definierten Signifikanzniveau ermöglichen sie bei konfirmatorischen Studien eine Entscheidung über Verwerfung oder Beibehaltung einer vorab formulierten Nullhypothese. Aussagen über die Effektstärke sind auf Grund von p-Werten nur sehr eingeschränkt möglich. Konfidenzintervalle liefern einen ausreichend plausiblen Bereich für den wahren Wert auf der Messebene des Punktschätzers. Aussagen zu Effektrichtung und -stärke sowie zum Vorliegen eines statistisch signifikanten Ergebnisses sind möglich. Abschließend ist festzustellen, dass es sich bei p-Werten und Konfidenzintervallen nicht um gegenläufige statistische Konzepte handelt. Bei Kenntnis der Stichprobengröße und der Streuung oder des Punktschätzers lassen sich aus p-Werten Konfidenzintervalle berechnen, und umgekehrt. Beide statistischen Konzepte ergänzen sich.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt im Sinne der Richtlinien des International Committee of Medical Journal Editors besteht.

Manuskriptdaten
eingereicht: 23. 7. 2008, revidierte Fassung angenommen: 21. 8. 2008

LITERATUR

1. Bland M, Peacock J: Interpreting statistics with confidence. The Obstetrician and Gynaecologist 2002; 4: 176–80.
2. Houle TT: Importance of effect sizes for the accumulation of knowledge. Anesthesiology 2007; 106: 415–7.
3. Faller, H: Signifikanz, Effektstärke und Konfidenzintervall. Rehabilitation 2004; 43: 174–8.
4. Greenfield ML, Kuhn JE, Wojtys EM: A statistics primer. Confidence intervals. AmJ Sports Med 1998; 26: 145–9. No abstract available. Erratum in: Am J Sports Med 1999; 27: 544.
5. Bender R, Lange St: Was ist ein Konfidenzintervall? Dtsch Med Wschr 2001; 126: 41.
6. Altman DG: Confidence intervals in practice. In: Altman DG, Machin D, Bryant TN, Gardner MJ. BMJ Books 2002; 6–9.
7. Weiss C: Intervallschätzungen. Die Bedeutung eines Konfidenzintervalls: In: Weiß C: Basiswissen Medizinische Statistik. Springer Verlag 1999; 191–2.
8. Moher D, Schulz KF, Altman DG für die CONSORT Gruppe: Das COSORT Statement: Überarbeitete Empfehlungen zur Qualitätsverbesserung von Reports randomisierter Studien im Parallel-Design. Dtsch Med Wschr 2004; 129: 16–20.
9. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF: Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999; 354: 1896–900.
10. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR: Publication bias in clinical research. Lancet 1991; 337: 867–72.
11. Shakespeare TP, Gebski VJ, Veness MJ, Simes J: Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and riskbenefit contours. Lancet 2001; 357: 1349–53. Review.
12. Gardner MJ, Altman DG: Confidence intervals rather than P-values: estimation rather than hypothesis testing. Br Med J 1986; 292: 746–50.
13. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S: Basic statistics for clinicians: 1. hypothesis testing. CMAJ 1995; 152: 27–32. Review.
14. ICH 9: Statisticlal Principles for Clinical Trials. London UK: International Conference on Harmonization 1998; Adopted by CPMP July 1998 (CPMP/ICH/363/96)
15. Feinstein AR: P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 1998; 51: 355–60.
16. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S: Basic statistics for clinicians: 2. interpreting study results: confidence intervals. CMAJ 1995; 152: 169–73.
17. Sim J, Reid N: Statistical inference by confidence intervals: issues of interpretation and utilization. Phys Ther 1999; 79: 186–95.
18. Gardner MJ, Altman DG. Confidence intervals rather than P values. In: Altman DG, Machin D, Bryant TN, Gardner MJ: Statistics with confidence. Confidence intervals and statistical guidelines. Second Edition. BMJ Books 2002; 15–27.

Anschrift für die Verfasser
Dr. med. Jean-Baptist du Prel, MPH
Zentrum für Kinder- und Jugendmedizin
Zentrum Präventive Pädiatrie Mainz
Langenbeckstraße 1
55101 Mainz
E-Mail: duprel@zpp.klinik.uni-mainz.de

SUMMARY

Confidence Interval or P-Value? Part 4 of a Series on Evaluation of Scientific Publications

Introduction: An understanding of p-values and confidence intervals is necessary for the evaluation of scientific articles. This article will inform the reader of the meaning and interpretation of these two statistical concepts.

Methods: The uses of these two statistical concepts and the differences between them are discussed on the basis of a selective literature search concerning the methods employed in scientific articles.

Results/Discussion: P-values in scientific studies are used to determine whether a null hypothesis formulated before the performance of the study is to be accepted or rejected. In exploratory studies, p-values enable the recognition of any statistically noteworthy findings. Confidence intervals provide information about a range in which the true value lies with a certain degree of probability, as well as about the direction and strength of the demonstrated effect. This enables conclusions to be drawn about the statistical plausibility and clinical relevance of the study findings. It is often useful for both statistical measures to be reported in scientific articles, because they provide complementary types of information.

Dtsch Arztebl 2009; 106(19): 335–9
DOI: 10.3238/arztebl.2009.0335

Key words: publications, clinical research, p-value, statistics, confidence interval

The English version of this article is available online:
www.aerzteblatt-international.de