Journal für
Hypertonie
Austrian Journal of Hypertension
Österreichische Zeitschrift für Hochdruckerkrankungen

Statistik für Kliniker – Was Mediziner von Statistik wissen sollten

Kvas E

Journal für Hypertonie - Austrian Journal of Hypertension 2014; 18 (3), 104-106

Offizielles Organ der Österreichischen Gesellschaft für Hypertensiologie

Österreichische Gesellschaft für Hypertensiologie
www.hochdruckliga.at

Indexed in EMBASE/Scopus

Krause & Pachernegg GmbH • Verlag für Medizin und Wirtschaft • A-3003 Gablitz
P.b.b. 02Z031106M, Verlagsort: 3003 Gablitz, Linzerstraße 177A/21 Preis: EUR 10,–

boso TM-2450

kleiner
leichter
leiser*

BOSCH +SOHN boso

Präzises ABDM – das neue 24-Stunden-Blutdruckmessgerät

Noch mehr Komfort für Ihre Patienten, noch mehr Leistungsfähigkeit für Sie.

- Kommunikation mit allen gängigen Praxis-Systemen über GDT
- Inklusive neuer intuitiver PC-Software profil-manager XD 6.0 für den optimalen Ablauf in Praxis und Klinik
- Übersichtliche Darstellung aller ABDM-Daten inklusive Pulsdruck und MBPS (morgendlicher Blutdruckanstieg)
- Gerät über eindeutige Patientennummer initialisierbar
- Möglichkeit zur Anzeige von Fehlmessungen (Artefakten)
- Hotline-Service

*im Vergleich mit dem Vorgängermodell boso TM-2430 PC 2

KLINISCH VALIDIERT
DIN
EN ISO 81060-2
CLINICALLY VALIDATED

Ausführliche Informationen erhalten Sie unter boso.at

boso TM-2450 | Medizinprodukt
BOSCH + SOHN GmbH & Co. KG
Handelskai 94-96 | 1200 Wien

Statistik für Kliniker – Was Mediziner von Statistik wissen sollten

E. Kvas

1. Irgendwann wird jeder Unterschied statistisch signifikant

In der klinischen Forschung vergleicht ein statistischer Test häufig zwei Gruppen miteinander, z. B. Therapiegruppe versus Placebogruppe. Nehmen wir als Parameter die Responderrate. Wir wollen also wissen, ob der Anteil von Patienten, welche auf die Behandlung ansprechen (Responder), in den beiden Gruppen verschieden groß ist.

1. Frage: Was heißt „verschieden"?

Auch wenn wir zwei Gruppen derselben Therapie miteinander vergleichen, wird der Unterschied im Response nicht genau 0 % sein. Es gibt immer einen Unterschied. Die Frage ist, ob der Unterschied zufällig beobachtet wurde oder ob es ihn höchstwahrscheinlich wirklich gibt. „Höchstwahrscheinlich wirklich" soll heißen, dass der in den Studiengruppen beobachtete Unterschied auch in den Populationen existiert, aus denen diese Gruppen gezogen wurden. Die Studiengruppen (Therapie und Placebo) sind Stichproben aus der Therapiepopulation und aus der Placebopopulation.

Bei kleinen Stichproben ist der beobachtete Unterschied umso eher auch zwischen den Populationen zu erwarten (also nicht zufällig), je größer dieser beobachtete Unterschied ist. Je kleiner der beobachtete Unterschied ist, desto wahrscheinlicher wurde dieser Unterschied bei kleinen Stichproben zufällig beobachtet.

Wenn wir immer größere Stichproben ziehen, so wird die Stichprobe der Population immer ähnlicher und ebenso wird der beobachtete Unterschied immer weniger zufällig, sondern in der Population wirklich vorhanden sein.

Der statistische Test berechnet für uns die Wahrscheinlichkeit, mit der mindestens der beobachtete Unterschied zwischen unseren Stichproben auftritt, wenn dieser Unterschied zwischen den Populationen gar nicht vorhanden ist. Das ist die Wahrscheinlichkeit von zumindest dem Studienergebnis, wenn die **Nullhypothese** (Therapie = Placebo) wahr ist. Diese Wahrscheinlichkeit bezeichnen wir als **p-Wert**. Wir entscheiden nun aufgrund des p-Werts, ob der beobachtete Unterschied eher zufällig oder doch systematischer Natur ist.

Tabelle 1: Anteil Responder in Gruppe 1 und Gruppe 2

Gruppe	Responder	Non-Responder	Summe
1	80	20	100
2	70	30	100

Tabelle 2: Anteil Responder in Gruppe 1 und Gruppe 2

Gruppe	Responder	Non-Responder	Summe
1	800	200	1000
2	700	300	1000

2. Frage: Was heißt „zufällig"?

Wir wollen nicht auf zufällige Unterschiede hereinfallen und falsche Entscheidungen treffen. Daher suchen wir uns eine Grenze für die Wahrscheinlichkeit einer falschen Entscheidung. Wir benutzen dafür den **Fehler 1. Art (alpha-Fehler).** Dieser Fehler beschreibt die Wahrscheinlichkeit eines Mindestunterschieds zweier Stichproben, wenn sich die Populationen nicht unterscheiden. Das ist genau der p-Wert des statistischen Tests. Nun ziehen wir eine Obergrenze ein. Meist sind das 5 %, das sogenannte **Signifikanzniveau alpha**. Die Entscheidungsregel lautet nun: Erzeugt der statistische Test beim Stichprobenvergleich einen **p-Wert** ≤ **alpha**, so sagen wir: Es gibt einen statistisch signifikanten Unterschied zwischen den Gruppen (Therapie vs. Placebo) und wir schließen daraus, dass dieser Unterschied auch in den Populationen vorhanden ist.

3. Frage: Was ist die Wahrheit?

Das bedeutet jedoch noch nicht, dass der Unterschied tatsächlich vorhanden ist. Wir haben nur gezeigt, dass solche Studienergebnisse im Gruppenvergleich für idente Populationen mit sehr geringer Wahrscheinlichkeit (p < 5 %) auftreten. Ausgeschlossen ist es nicht. Auch wenn wir uns bei einem p-Wert < 0,0001 % für einen Unterschied zwischen den Populationen entscheiden, kann diese Entscheidung noch immer falsch sein. Die Irrtumswahrscheinlichkeit wird niemals Null, sondern nur sehr klein.

Beispiel: Tabelle 1 zeigt den Anteil an Respondern in Gruppe 1 (80 %) und Gruppe 2 (70 %).

Der Chi²-Test ergibt einen p-Wert = 0,1412. Unterschiede von mindestens 10 Prozentpunkten (80–70 %) im Response können also mit einer Wahrscheinlichkeit von 14,12 % auftreten, wenn Gruppe 1 und Gruppe 2 aus der selben Population stammen (d.h. wahrer Gruppenunterschied = 0). Wir entscheiden also: kein statistisch signifikanter Unterschied zwischen den Gruppen (p > alpha = 5 %).

In Tabelle 2 wurde die Stichprobe verzehnfacht. Der Unterschied zwischen den Gruppen (der Effekt) ist nach wie vor 10 Prozentpunkte. Der Chi²-Test liefert jetzt einen p-Wert = 0,0000003. Wir beobachten also einen statistisch „hoch signifikanten" Unterschied (p < alpha = 5 %).

4. Frage: Was ist geschehen?

In Tabelle 2 haben wir in den großen Stichproben wesentlich mehr Information aus den Populationen (die Stichproben sind den Populationen ähnlicher) als in Tabelle 1. Entsprechend stammt der beobachtete Unterschied in Tabelle 2 sehr viel we-

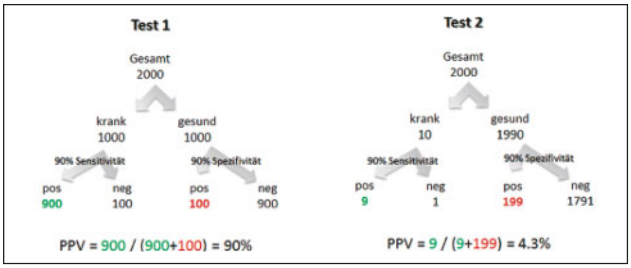

Abbildung 1: PPV als natürliche Häufigkeiten

niger wahrscheinlich aus identen Populationen als in Tabelle 1. Kleine Stichproben wie in Tabelle 1 können mit größerer Wahrscheinlichkeit zufällige Unterschiede zeigen, wenn die Stichproben aus identen Populationen stammen.

5. Frage: Was folgt daraus?

Je größer die Stichproben, desto eher wird das Testergebnis statistisch signifikant sein. Ganz egal, wie klein der beobachtete Unterschied ist, irgendwann wird er statistisch signifikant. Denn für die Statistik spielt klinische Relevanz keine Rolle.

Conclusio: Statistische Signifikanz ≠ Klinische Relevanz

2. Die Aussagekraft einer statistisch signifikanten klinischen Studie geht (momentan) gegen Null

Wir wissen nun, was es bedeutet, wenn ein Studienergebnis einen signifikanten Unterschied zwischen zwei Gruppen gezeigt hat: Wenn die Stichproben aus derselben Population stammen, ist es höchst unwahrscheinlich (p-Wert), dass in der Studie große Unterschiede beobachtet werden (Aussage 1).

Aber bedeutet das auch: Wenn die Studie ein statistisch signifikantes Ergebnis gezeigt hat, so stammen die Stichproben höchst wahrscheinlich aus unterschiedlichen Populationen (Aussage 2).

Das hört sich sehr ähnlich an. Wo ist der Unterschied zwischen den beiden Aussagen? Das eine Mal setzen wir voraus, dass die sogenannte **Nullhypothese** (z. B. „Therapie ist gleich gut wie Placebo“) wahr ist. Und unter dieser Bedingung schätzen wir den p-Wert, also die Wahrscheinlichkeit für einen Unterschied, wenn dieser gar nicht existiert (Aussage 1).

Bei Aussage 2 setzen wir voraus, dass die Studie ein statistisch signifikantes Ergebnis gezeigt hat. Unter dieser Bedingung (signifikantes Ergebnis) schätzen wir die Wahrscheinlichkeit, dass die Studienhypothese (z. B. „Therapie ist besser als Placebo“) wahr ist. Diese Wahrscheinlichkeit bezeichnet man als **positiven Vorhersagewert** der Studie oder „positive predictive value“ (PPV).

Der PPV stammt aus der Epidemiologie und beschreibt die Wahrscheinlichkeit, mit der ein Patient krank ist, wenn sein Befund des diagnostischen Tests positiv ist (z.B. Mamma-Screening, HIV-Test, Pap-Abstrich etc.). Solche Tests sind mit einer bestimmten Wahrscheinlichkeit richtig positiv, wenn der Patient krank ist (Sensitivität), und richtig negativ, wenn der

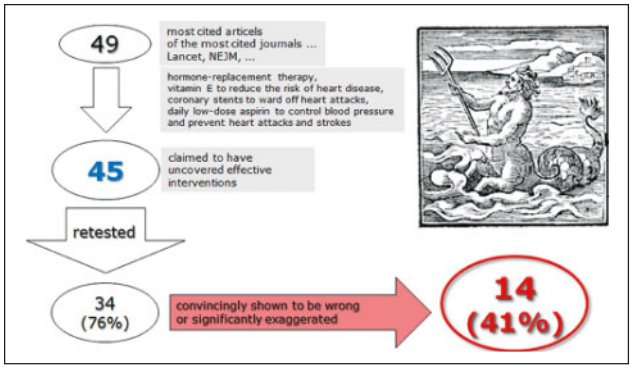

Abbildung 2: Ioannidis‘ Untersuchung der besten der besten medizinischen Publikationen

Patient gesund ist (Spezifität). Andersherum ist der Patient mit einer bestimmten Wahrscheinlichkeit krank, wenn der Test positiv ist (PPV) bzw. gesund, wenn der Test negativ ist (negative predictive value, NPV).

Abbildung 1 zeigt den PPV mithilfe von natürlichen Häufigkeiten, welche nach Gerd Gigerenzer [1] leichter zu verstehen sind als Häufigkeitstabellen und die zugehörigen bedingten Wahrscheinlichkeiten. Wir sehen zwei diagnostische Tests. Beide Tests untersuchen 2.000 Patienten und können mit 90 % Wahrscheinlichkeit einen kranken Patienten richtig positiv befunden (Sensitivität) und ebenso mit 90 % Wahrscheinlichkeit einen gesunden Patienten richtig negativ bewerten (Spezifität). Dennoch liegt der PPV für Test 1 mit 90 % sehr hoch und für Test 2 mit 4,3 % sehr niedrig.

Der entscheidende Faktor ist die **Prävalenz**, also die Wahrscheinlichkeit, mit der die Erkrankung in der Bevölkerung überhaupt auftritt. Ist die Erkrankung in der Bevölkerung häufig (Test 1, Prävalenz = 50 %), so ist auch die Wahrscheinlichkeit hoch, dass ein positiver Befund richtig ist. Kommt die Erkrankung in der Bevölkerung sehr selten vor (Test 2, Prävalenz = 0,5 %), so ist die Wahrscheinlichkeit dafür, dass ein positiver Befund richtig ist, sehr gering (PPV = 4,3 %).

Mit diesem Ansatz kann man sich der Frage nähern, wie wahrscheinlich eine Studienhypothese wahr ist, wenn die Studie ein signifikantes Ergebnis gezeigt hat (PPV). Auf humorvolle Weise setzen sich Hans-Peter Beck-Bornholdt und Hans-Hermann Dubben in ihrem Buch „Der Schein der Weisen“ mit dem Thema auseinander [2].

In seinem aufsehenerregenden Aufsatz aus dem Jahr 2005 hat John Ioannidis zu erklären versucht, warum die meisten publizierten Studienergebnisse falsch sind [3]. Aus davor durchgeführten Studien ist er zu dem Schluss gekommen, dass in der wissenschaftlichen Fachliteratur die Prävalenz publizierter, wahrer Studienhypothesen sehr niedrig ist. Das heißt: Die Anzahl an Studien, welche einen Effekt untersuchen, der wirklich existiert, ist sehr gering. Eine dieser Untersuchungen [4] (Abb. 2) beschäftigte sich mit den 49 meistzitierten Publikationen (> 1000 citations) der meistzitierten Journale (NEJM, JAMA, Lancet) der medizinischen Fachliteratur der Jahre 1990 bis 2003, von welchen 45 Publikationen einen signifikanten Effekt beschrieben. Unter diesen publizierten Studien finden sich Namen wie HOPE, WOSCOPS,

Tabelle 3: PPV einer signifikanten Studie in Abhängigkeit von alpha, Power und *a priori*-Chance einer wahren Studienhypothese

alpha (%)	Power (%)	Odds (wahr:falsch)	PPV (%)
5,0	80,0	2:1	97
5,0	80,0	1:2	89
5,0	80,0	1:10	62
5,0	80,0	1:100	14
5,0	80,0	1:10.000	0,16

Tabelle 4: PPV einer signifikanten Studie in Abhängigkeit von alpha, Power und *a priori*-Chance einer wahren Studienhypothese bei drastisch gesenktem Signifikanzniveau

alpha (%)	Power (%)	Odds (wahr:falsch)	PPV (%)
0,1	80,0	2:1	100
0,1	80,0	1:2	100
0,1	80,0	1:10	99
0,1	80,0	1:100	89
0,1	80,0	1:10.000	7.4

CAPRIE, UKPDS34, LIPID etc. Von diesen 45 Studien wurden 34 (76 %) mit größeren Stichproben oder höherer Qualität (Randomisierung) wiederholt. Die restlichen 11 (24 %) Studienergebnisse wurden nie einer Überprüfung zugeführt. Von den 34 Studien konnten 20 (59 %) Studienergebnisse in den nachfolgenden Kontrollstudien bestätigt werden. 14 (41 %) Studienergebnisse wurden in den Wiederholungsstudien als falsch (n = 7) oder mit signifikant überschätztem Effekt (n = 7) bewertet. Dies ist einer der Gründe, warum Ioannidis der medizinischen Forschung das *Proteus*-Phänomen zuordnet: *Proteus* als Gott der griechischen Mythologie steht für die Wandlungsfähigkeit (Metamorphose).

Die oben genannten Ergebnisse beziehen sich nur auf die besten Publikationen der besten Journale. Einer groben Schätzung zufolge wächst die Literaturdatenbank der Medline (PubMed) jährlich um mehr als 1 Million (n = 1,000.000) Neupublikationen. Unterschiedliche Forschungsqualität in verschiedenen Fachgebieten ergibt unterschiedliche Prävalenzen von *a priori* wahren untersuchten Studienhypothesen. Diese Prävalenzen kann man als *a priori*-Chance (Odds) abschätzen: die Chance, dass eine wahre Hypothese durch eine Studie untersucht wird.

Ob ein Fachgebiet qualitativ gut beforscht wird, hängt von verschiedenen Parametern ab, welche das Studienergebnis verzerren können (Größe der Studien, Größe der Effekte, Sponsoring, Hype, ...). In exzellent beforschten Gebieten wurde diese Chance (wahr : falsch) von Ioannidis mit 2:1 bis 1:2 eingeschätzt. Die Realität der Literaturdatenbanken führt uns aber eher zu Odds von 1:10 oder 1:100.

Tabelle 3 zeigt deutlich, dass die Aussagekraft (PPV) einer signifikanten Studie nur in exzellenten Forschungsgebieten brauchbare Werte annimmt. Odds von 1:10.000 werden Genscreenings zugeordnet, welche in einer Studie unter 100.000 Genen die vielleicht 10 relevanten Gene auf mögliche Krankheitshäufungen untersuchen (10:100.000). Dabei wird für jedes Gen getestet, ob der Genpolymorphismus signifikant mit der Krankheitshäufigkeit zusammenhängt. Aus statistischer Sicht wäre es überraschend, wenn es bei solchen Untersuchungen nicht zu vielen signifikanten Ergebnissen kommen würde.

Nachdem es kaum eine Möglichkeit gibt, die Forscher zu zwingen, ihre Forschungstätigkeit auf ein deutlich höheres Qualitätsniveau zu heben, schlägt Ioannidis ein Senken des Signifikanzniveaus vor.

Tabelle 4 zeigt den PPV einer signifikanten Studie, wenn man das Signifikanzniveau deutlich auf alpha = 0,1% senkt. Nun gewinnen auch Studienergebnisse an Bedeutung, die aus Forschungsgebieten von mittelmäßiger Qualität stammen. Für die Genforschung bringt aber auch dieser Ansatz nichts. Hier wird man wohl neue Methoden entwickeln müssen, bevor der genetische Fingerprint flächendeckend zum Einsatz kommt.

Literatur:

1. Gigerenzer G. Risiko: Wie man die richtigen Entscheidungen trifft. C. Bertelsmann Verlag, 2013.

2. Beck-Bornholdt HP, Dubben HH. Der Schein der Weisen. Rowohlt TB, 2003.

3. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005; 2: e124, 696–701.

4. Ioannidis JPA. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. JAMA 2005; 294: 218–28.

Korrespondenzadresse:
DI Erich Kvas
Hermesoft
A-8046 Stattegg, Forstweg 96b
E-Mail: Office@Hermesoft.at
www.hermesoft.at

Mitteilungen aus der Redaktion

Abo-Aktion

Wenn Sie Arzt sind, in Ausbildung zu einem ärztlichen Beruf, oder im Gesundheitsbereich tätig, haben Sie die Möglichkeit, die elektronische Ausgabe dieser Zeitschrift kostenlos zu beziehen.

Die Lieferung umfasst 4–6 Ausgaben pro Jahr zzgl. allfälliger Sonderhefte.

Das e-Journal steht als PDF-Datei (ca. 5–10 MB) zur Verfügung und ist auf den meisten der marktüblichen e-Book-Readern, Tablets sowie auf iPad funktionsfähig.

☑ **Bestellung kostenloses e-Journal-Abo**

Besuchen Sie unsere

zeitschriftenübergreifende Datenbank

☑ **Bilddatenbank** ☑ **Artikeldatenbank** ☑ **Fallberichte**