AUS DEM DEUTSCHEN NETZWERK EVIDENZBASIERTE MEDIZIN

„Statistisch signifikant“ ist nicht genug

Der p-Wert allein beantwortet nicht die wichtige Frage nach Zufall und Replizierbarkeit

VON HANS-HERMANN DUBBEN IM AUFTRAG DES DEUTSCHEN NETZWERKS EVIDENZBASIERTE MEDIZIN E.V. (DNEBM – WWW.EBM-NETZWERK.DE)

In der medizinischen Forschung sind statistische Signifikanztests sehr verbreitet. 16 Prozent aller Abstracts und 55 Prozent aller *full papers* in *Pubmed* berichten p-Werte. In 96 Prozent dieser Publikationen gibt es mindestens ein „statistisch signifikantes“ Ergebnis. Heißt das, dass in 96 von 100 Studien etwas Wichtiges herauskommt? Schön wär's! Es ist eher ein Indiz, dass im Wissenschaftsbetrieb etwas nicht stimmt. In einem Review [1] wurde gezeigt, dass bis zu 89 Prozent von Studien mit „statistisch signifikantem“ Ergebnis nicht replizierbar waren. Diese *replication crisis* ist ein großes ethisches, wissenschaftliches und wirtschaftliches Problem.

Das Wort „signifikant“ heißt in etwa „bedeutsam“. Das Ergebnis einer Studie, egal ob positiv oder negativ, ist aber nur bedeutsam, wenn die Fragestellung bedeutsam war. Diese Bedeutsamkeit muss der klinische Forscher *vor* der Studie belegen. Die Deklaration von Helsinki verpflichtet ihn dazu. Sie untersagt bedeutungslose Experimente am Menschen.

Das Attribut „statistisch signifikant“ heißt lediglich „aus statistischer Sicht bedeutsam“. Für wen und inwiefern das Ergebnis *relevant* ist, steht auf einem ganz anderen Blatt.

Angenommen, ein Forscher untersucht, welche der Therapien A und B die bessere ist (zum Beispiel welche die höhere Ansprechrate hat); und die Patienten mit Therapie A zeigen das bessere Ergebnis. Daraus kann er nicht schließen, dass Therapie A grundsätzlich besser ist. Er muss noch ein paar Fragen mehr abwägen: Kann das Ergebnis Zufall sein? Inwieweit kann er systematische Fehler seiner Studie ausschließen? Was spricht dafür und was spricht dagegen, dass Therapie A die *Ursache* für das bessere Ergebnis ist? Erst danach kann der Forscher mit größerer Sicherheit, aber niemals mit Gewissheit, die Wirksamkeit einschätzen.

KANN ES ZUFALL SEIN?

Replizieren – auch Reproduzieren oder schlicht Wiederholen genannt – ist ein wichtiges Merkmal von Wissenschaftlichkeit und entscheidendes Abgrenzungsmerkmal gegen Pseudowissenschaften. Ob ein Experiment replizierbar ist, erfährt man streng genommen nur durch replizieren. Statistische Signifikanztests sind der Versuch, den Daten eines Experimentes oder einer Studie zu entlocken, mit welcher Wahrscheinlichkeit sie auf Zufall beruhen könnten.

Die Ergebnisse der obigen Studie mögen 70% für A und 60% für B lauten. Mit der Annahme, dass Therapie A und B eigentlich gleich wirksam sind (das ist die altbekannte Nullhypothese), kann der Statistiker die Wahrscheinlichkeit ausrechnen, mit der das Ergebnis 70% : 60% zufällig auftritt. Diese Wahrscheinlichkeit

heißt „p-Wert". Sie ist ausschließlich eine Aussage über das Ergebnis. Der p-Wert ist nicht die Wahrscheinlichkeit, mit der die Nullhypothese wahr ist. Um einen weitreichenden Trugschluss aufzuzeigen, dasselbe noch einmal in Kurzform:

A.) Wenn die Nullhypothese wahr ist, dann beträgt die Wahrscheinlichkeit (der p-Wert), dass das Ergebnis „70% : 60%" zufällig eintritt, 4%. (Wie die 4% berechnet wurden ist kein Geheimnis, soll uns jetzt aber nicht interessieren.)

Viele Forscher folgern daraus:

B.) Wenn das Ergebnis „70% : 60%" tatsächlich eingetreten ist, dann beträgt die Wahrscheinlichkeit, dass die Nullhypothese wahr ist, 4%. (Und weiter: Die Wahrscheinlichkeit, dass die Nullhypothese nicht wahr ist, und Therapie A und B unterschiedlich wirksam sind, beträgt 100% - 4% = 96%.)

Dieses Argument ist falsch. In einem Kontext, in dem wir uns alle sicherer fühlen, fällt das schnell auf:

A.) Wenn uns ein Mensch begegnet, dann beträgt die Wahrscheinlichkeit, dass es zufällig ein Mann ist, 50%.

Wenn die obige Folgerung vieler Forscher richtig ist, dann müsste auch diese richtig sein:

B.) Wenn wir tatsächlich einem Mann begegnen, dann beträgt die Wahrscheinlichkeit, dass er ein Mensch ist, 50%.

Die zweite Argumentation ist analog zur ersten. Beide sind falsch. Das ist lange bekannt [2]. Es ist keine Haarspalterei, auf diese logische Fehlleistung hinzuweisen. Sie hat weitreichende Konsequenzen [3-8] und trägt maßgeblich zur *replication crisis* bei.

Ein weiterer logischer Fehler ist es, Ergebnisse für replizierbar zu halten, nur weil der p-Wert kleiner ist als beispielsweise 5%, zumal dieses so genannte „Signifikanzniveau" völlig willkürlich gewählt ist [9].

Der p-Wert ist relativ einfach zu berechnen und lässt Datenanalysen besonders wissenschaftlich und mathematisch exakt erscheinen. Das könnte die Beliebtheit des p-Wert-Rituals erklären. Der p-Wert allein beantwortet die wichtige Frage nach Zufall und Replizierbarkeit nicht [10]. Dazu müssen weitere Größen wie die *Power* der Studie und die Vortestwahrscheinlichkeit in Betracht gezogen werden [2, 5, 7, 8, 11].

INWIEWEIT KANN EIN SYSTEMATISCHER FEHLER DER STUDIE AUSGESCHLOSSEN WERDEN?

Nehmen wir mal an, das Ergebnis sei nach reiflicher Überlegung sehr wahrscheinlich nicht zufällig. Dann besteht Hoffnung, dass es tatsächlich auf den getesteten Therapien beruht. Trotzdem sollte man seine Daten – oder die Publikation, die man gerade liest – auf systematische Fehler abklopfen. Schließlich ist die Auswahl groß. Es gibt über 235 Arten von Bias, Fehlern und Verzerrungen in Studien [12]. Welche können in einer bestimmten Studie eine Rolle spielen? Welche kann man ausschließen, weil die Studienleitung belegbar wirksame Gegenmaßnahmen ergriffen hatte? Besteht die Möglichkeit der Manipulation?

„Statistische Signifikanz" verrät uns nicht die *Ursache* für die besseren Ergebnisse in Gruppe A. Am Ende einer Studie wissen wir zunächst nur, dass in der einen Gruppe *irgendetwas* zu besseren Resultaten führte.

WAS SPRICHT DAFÜR, DASS THERAPIE A DIE URSACHE FÜR DAS BESSERE ERGEBNIS IST?

Was spricht für eine Kausalbeziehung zwischen der Therapie und dem Ergebnis? Bei dieser Frage ist eine Liste von Kriterien aus der epidemiologischen Forschung [13] sehr hilfreich. Trotz der (endlosen) philosophischen Tiefen des Begriffs Kausalität wird pragmatisch gefragt: Wenn Ereignis E (Auftreten einer Krankheit oder Heilung von einer Krankheit) und Merkmal M (Umweltfaktor oder eine Intervention) miteinander assoziiert sind oder korrelieren – unter welchen

Umständen dürfen wir dann an einen Kausalzusammenhang glauben (jedenfalls solange uns keine neuen Erkenntnisse wieder zweifeln lassen)?

Folgendes sollte bedacht werden [13]: Wie groß ist der Effekt? Wurde das Ergebnis von anderen Forschergruppen, an anderen Orten und auch unter anderen Rahmenbedingungen repliziert? Gibt es eine Dosis-Effekt-Beziehung? Gibt es eine plausible, beispielsweise biologische Erklärung für den mutmaßlichen Kausalzusammenhang? Wenn ja, dann ist das sehr ermutigend. Wenn nein, dann sollte man das nicht überschätzen, denn Plausibilität hängt vom jeweiligen zeitgenössischen Wissen ab. Passen die Erkenntnisse aus Experimenten mit Zellen, Mäusen, Menschen widerspruchsfrei zusammen? Der mutmaßliche Kausalzusammenhang und Wirkmechanismus sollte nicht in krassem Widerspruch zu validen naturwissenschaftlichen Erkenntnissen stehen. Verändert der Entzug der mutmaßlichen Ursache das Auftreten der Wirkung? Ist für einen ähnlichen, analogen Zusammenhang eine Ursache-Wirkung-Beziehung bekannt?

Dies ist keine Checkliste, sondern lediglich eine Hilfe zum strukturierten Nachdenken. Je mehr dieser Kriterien für Kausalität sprechen, umso größer ist die Vortestwahrscheinlichkeit, dass in obiger Studie tatsächlich Therapie A die Ursache für das bessere Resultat ist.

Die Frage, ob eine Korrelation kausal ist, lässt sich nicht mit Statistik allein beantworten. Auch kann nicht mit einer einzelnen Studie über einen Kausalzusammenhang entschieden werden. Vielmehr ist eine Gesamtschau von bekannten Zusammenhängen, Ergebnissen der Grundlagenforschung und der klinischen Forschung notwendig.

Das geht natürlich weit über das hinaus, was für den Einzelnen neben Praxis- und/oder Klinikbetrieb leistbar ist. Hier können systematische Reviews, Metaanalysen und Leitlinien hilfreich sein, in denen nicht nur Zahlen zusammengetragen und p-Werte aufgelistet werden, sondern auch eine sorgfältige Gesamtschau und Bewertung der Evidenz im obigen Sinne durchgeführt wird.

ZUSAMMENFASSUNG

Statistische Signifikanz darf nicht mit Relevanz verwechselt werden. Wissenschaftliche Fragen werden nicht durch einzelne Studien geklärt. Statistische Signifikanztests allein sagen sehr wenig über die Qualität einer Studie und deren Replizierbarkeit aus. Es muss immer erwogen werden, ob Studienergebnisse auf einem systematischen Fehler beruhen könnten. Es muss kritisch erwogen werden, welche Argumente für beziehungsweise gegen einen Kausalzusammenhang von Intervention und Ergebnis sprechen. ■

PD Dr. rer. nat. Hans-Hermann Dubben
Institut und Poliklinik für Allgemeinmedizin
Universitätsklinikum Hamburg-Eppendorf
Martinistraße 52, 20246 Hamburg
Tel 040 7410-56064

1.) Freedman LP, Cockburn IM, Simcoe TS: The Economics of Reproducibility in Preclinical Research. PLoS Biol 2015; 13(6): e1002165. doi:10.1371/journal.pbio.1002165
2.) Bayes T: An Essay towards Solving a Problem in the Doctrine of Chances. Phil. Trans. 1764; 53: 370-418.
3.) Beck-Bornholdt HP, Dubben HH: Is the pope an alien? Nature 1996; 381: 730.
4.) Beck-Bornholdt HP und Dubben HH: Der Schein der Weisen - Irrtümer und Fehlurteile im täglichen Denken. Rowohlt Verlag, 2003.
5.) Ioannidis JPA: Why Most Published Research Findings Are False. PLoS Med 2005; 2(8): e124.
6.) Nature Editorial: Number crunch. Nature 2014; 506: 131.
7.) Nuzzo R: Statistical errors. Nature 2014; 506: 150-152.
8.) Nuzzo R: Wenn Forscher durch den Signifikanztest fallen. Spektrum der Wissenschaft 2014; http://www.spektrum.de/news/wenn-forscher-durch-den-signifikanztest-fallen/1224727
9.) Sterne JAC und Smith GD: Sifting the evidence — what's wrong with significance tests? British Medical Journal 2001; 322, 226-231.
10.) Goodman SN: Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann Intern Med. 1999; 130: 995-1004.
11.) Dubben HH, Beck-Bornholdt HP: Die Bedeutung der statistischen Signifikanz. Kölner Zeitschrift für Soziologie und Sozialpsychologie 2004; Sonderheft 44: 61-74.
12.) Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. J Clin Epidemiol 2010; 63: 1205–15.
13.) Bradford-Hill A, 1965: The Environment and Disease: Association or Causation? Proc R Soc Med. 1965; 58: 295–300.

EbM Netzwerk