DER ARZNEIMITTELBRIEF

Jg. 55, S. 20DB01; Ausgabe 02 / 2021

Gehoert die statistische Signifikanz in den Ruhestand?

Vor ungefähr 100 Jahren hat Ronald A. Fisher wesentlich zur breiten Einführung des Konzepts der statistischen Signifikanz in die Wissenschaft beigetragen (1). Schon bald darauf begannen Forscher, Missbrauch und Fehlinterpretationen der statistischen Signifikanz und des p-Werts zu kritisieren (vgl. 2). Die Diskussion erreichte ihren Höhepunkt, als der weltweit größte Verbund von Statistikern, die American Statistical Association, sich im Jahr 2019 gegen die Verwendung des Begriffs der statistischen Signifikanz aussprach (3). Auch andere Begriffe wie „signifikant unterschiedlich“, „p 0,05" und "nicht signifikant" sollten nicht eingesetzt werden, egal ob in Worten, durch Sternchen in einer Tabelle oder auf andere Weise ausgedrückt. „Das Werkzeug wurde zum Tyrannen“, kritisieren die Wissenschaftler, da der aus ihrer Sicht vollkommen überbewertete p-Wert darüber entscheidet, ob Hypothesen bestätigt, Studien veröffentlicht und Produkte auf den Markt gebracht werden. Kurz darauf forderten auch Wissenschaftler in einem Beitrag in der Fachzeitschrift Nature, auf den Begriff der statistischen Signifikanz zu verzichten und p-Werte nicht mehr in „signifikant“ und „nicht-signifikant“ zu dichotomisieren (4). Mehr als 800 Forscherinnen und Forscher unterzeichneten den Aufruf. In der Folge hat beispielsweise das N. Engl. J. Med. neue Leitlinien veröffentlicht, nach denen p-Werte sparsam zu verwenden sind (5). Aktuell erläutern Autoren im Arzneimittelbulletin „Australian Prescriber“, der wie „DER ARZNEIMITTELBRIEF“ Mitglied in der International Society of Drug Bulletins ist, warum es an der Zeit ist, sich vom Konzept der statistischen Signifikanz zu lösen und p-Werte vorsichtig einzusetzen (6).

Der p-Wert (engl.: probability value oder p-value) gibt die Wahrscheinlichkeit unter Annahme einer Nullhypothese wieder, das festgestellte Prüfungsergebnis zu erhalten oder Ergebnisse, die noch stärker von der Nullhypothese abweichen (1). In der Medizin hat sich für die Nullhypothese durchgesetzt, dass Arzneimittel oder Interventionen keine oder „null“ Wirksamkeit haben. Der p-Wert quantifiziert die Verträglichkeit der Daten mit der Nullhypothese von 0 bis 1. Je kleiner der p-Wert, desto deutlicher spricht das beobachtete Ergebnis gegen die Nullhypothese. Statistische Signifikanz bedeutet also lediglich, dass Studienergebnisse ein willkürlich festgelegtes Niveau überschritten haben, an dem sie nicht mehr mit der Nullhypothese vereinbar sind. Es gibt jedoch viele Gründe, warum Ergebnisse nicht mit der Nullhypothese vereinbar sind: Beispielsweise, weil eine ungeeignete Teststatistik angewandt wurde, es zu einem Bias kam bei der Auswahl der Probanden für die Studie oder bei ihrer Nachbeobachtung (Selektionsbias) oder es Fehler bei der Erhebung der Daten gab (1). Statistische Signifikanz gibt außerdem keine Sicherheit darüber, wie sorgfältig eine Studie durchgeführt wurde. Ob eine klinische Studie eine statistische Signifikanz zeigen kann, hängt unter anderem von der Zahl der analysierten Patienten ab. Statistische Signifikanz allein bedeutet nicht, dass Ergebnisse richtig sind und die Nullhypothese falsch. Vor allem aber bedeutet statistische Signifikanz nicht, dass die Ergebnisse klinisch relevant sind.

Der p-Wert wird häufig missverstanden und beispielsweise als Wahrscheinlichkeit für die Richtigkeit der Nullhypothese missinterpretiert (1). Eine australische Studie hat untersucht, wie Kliniker einen p-Wert verstehen, wie er typischerweise in medizinischen Fachpublikationen angegeben wird (7). Das häufigste Missverständnis war, dass der p-Wert die numerische Wahrscheinlichkeit für ein Ereignis in der „real world" angibt. Doch beispielsweise bedeutet ein p-Wert von 0,05 nicht, dass die Nullhypothese „mit 95%-iger Sicherheit" falsch ist. Vielmehr beschreibt der p-Wert, wie wahrscheinlich die Ergebnisse unter einer angenommenen Hypothese sind. Dies scheint zunächst sehr ähnlich wie „Die Wahrscheinlichkeit für eine Hypothese bei den Ergebnissen". Die Bedeutung der Reihenfolge erklärt der Statistiker Stephen Senn mit folgendem Beispiel (8): „Ist der Papst katholisch? Die Antwort ist ja. Ist ein Katholik Papst? Die Antwort ist: Wahrscheinlich nicht".

Zu den weiteren Fehlinterpretationen gehörte, dass der p-Wert häufig anhand des Grenzwerts beurteilt wurde ($p = 0,05$). Dies führt jedoch zu einer inadäquaten Dichotomisierung der Forschungsergebnisse. Ein $p = 0,04$ ist nicht grundlegend unterschiedlich von einem $p = 0,06$.

Wichtiger als p-Wert und statistische Signifikanz ist die klinische Relevanz einer Studie. Um sie einzuschätzen, muss zunächst geprüft werden, ob die Fragestellung der Untersuchung plausibel ist und ob Patienten, Interventionen, Vergleichstherapie und Endpunkte zu Recht ausgewählt wurden. Dann muss die interne Validität beurteilt werden. Um interne Validität zu gewährleisten, müssen beispielsweise die Patientengruppen zu Beginn der Studie möglichst gleich sein, und die Behandlung darf sich nur in der speziell gewählten Intervention unterscheiden. Die Bewertung der Studienergebnisse sollte sich fokussieren auf den primären Endpunkt, die Größe des Effekts und die Genauigkeit, mit der sie geschätzt werden konnte. Dies wird in den Konfidenzintervallen angegeben. Auch hier ist Vorsicht geboten: Eine große Arzneimittelstudie mit Männern kann zu einer genauen Schätzung des Ergebnisses führen, für Frauen aber trotzdem nicht zutreffen (6).

Fazit: Statistische Signifikanz und p-Wert allein erlauben keine Bewertung davon, wie glaubwürdig, reproduzierbar, relevant oder publikationswürdig eine Untersuchung ist. Die Beurteilung von Studienergebnissen sollte sich konzentrieren auf die Schätzung der Effektgröße, ihre Genauigkeit und die klinische Signifikanz.

Literatur

1. Stang, A. und Kowall, B.: GMS Med. Inform. Biom. Epidemiol. 2020. Link zur Quelle
2. AMB 2014, **48**, 56DB01. Link zur Quelle
3. Wasserstein, R.L., et al.: The American Statistician 2019, **73 Sup1**, 1. Link zur Quelle
4. Amrhein, V., et al.: Nature 2019, **567**, 305. Link zur Quelle
5. Harrington, D., et al.: N. Engl. J. Med. 2019, **380**, 285. Link zur Quelle
6. Frank, O., et al.: Australian Prescriber 2021, **44**, 16. Link zur Quelle
7. Tam, C.W.M., et al.: AJGP 2018, **47**, 705. Link zur Quelle
8. Senn, S.: Significance 2013, **10**, 40. Link zur Quelle