C. Baulig[1], B. Al-Nawas[2], F. Krummenauer[1]

p-Werte – statistische Signifikanz ist keine klinische Relevanz!

Neben der ausführlichen Deskription der in einer Studie erhobenen Daten muss auch deren statistische Signifikanz beleuchtet werden, also die Absicherung der Übertragbarkeit von in einer Studie erhaltenen Ergebnisse auf andere (nicht in der Studie enthaltene) Patienten. Sehr vereinfacht formuliert beschreibt dabei ein p-Wert die statistische Wahrscheinlichkeit, mit welcher ein Studienergebnis unter Umständen „fälschlich" auf andere Patienten übertragen werden kann. Ein p-Wert < 5 % zum Vergleich zweier Therapieregimes besagt sehr vereinfacht, dass ein in der Studie aufgetretener Therapieunterschied „guten Gewissens" (statistisch abgesichert) auf andere Patienten übertragen werden kann, da höchstens mit einer Rest-Irrtumswahrscheinlichkeit von 5 % dabei eine falsche Aussage auf diese anderen Patienten projiziert wird. Der p-Wert sagt jedoch nichts über Ausmaß und klinischen Nutzen des beobachteten Unterschiedes zwischen den Therapien aus: Nicht selten wird jedoch die „statistische Signifikanz" – also die Übertragbarkeit – eines Studienergebnisses mit dessen „klinischer Relevanz" – also dem Nutzen für den Patienten – verwechselt: Neben der statistischen Signifikanz muss daher in jedem Fall auch die klinische Größenordnung des beobachteten Studienergebnisses (etwa durch Angabe des medianen Unterschiedes zwischen konkurrierenden Therapieregimes) beschrieben werden. Insofern beschreiben die Termini „statistisch signifikant" und „klinisch relevant" zwei verschiedene Ebenen eines Studienergebnisses und sollten stets beide in Ergebnisteil und Diskussion einer Studienpublikation einfließen.

Schlüsselwörter: p-Wert, statistische Signifikanz, klinische Relevanz

Primäre klinische Endpunkte

Wird etwa die Wirksamkeit eines Chlorhexidin-Chips zur adjunktiven adjuvanten antibakteriellen Therapie bei aggressiver Parodontitis überprüft, so ist es üblich einen randomisierten Vergleich mit einem Placebo oder mit einem etabilierten Standardpräparat (systemische Gabe von Amoxicillin/Metronidazol) durchzuführen [1]. Dabei muss vor Beginn der Studie festgelegt werden, anhand welcher Zielparameter die therapeutische Wirkung der Präparate erfasst werden: Soll die klinische Taschensondierungstiefe nach drei bzw. sechs Monaten betrachtet werden, oder interessiert ein radiologischer Vergleich der Attachmentlevel? Oder aber werden Entzündungsparameter und mikobiologische Taschenbesiedelungen zum Vergleich der Therapiemodi in den Vordergrungd gestellt? In jedem Fall muss vor Beginn der Studie explizit ein sogenannter primärer Endpunkt fixiert sein, also desjenigen Kriteriums, an welchem die Wirksamkeit des zu testenden Therapieansatzes und damit das Ergebnis der zentralen Studienfragestellung festgemacht werden sollen.

Natürlich werden in jeder Studie auch zahlreiche weitere Informationen erfasst. Diese *sekundären Endpunkte* werden ebenso gründlich dokumentiert und ausgewertet, wie der primäre. Ihre Auswertungsergebnisse haben zwar für die Kernfragestellung einer Studienpublikation nur sekundäre Wertigkeit, aber nicht selten hohe Relevanz für die Aufwerfung neuer Arbeitshypothesen z. B. zu unerwünschten Arzneimittelwirkungen, Patientensicherheit und Spezifika in der Wirksamkeit bei Subgruppen innerhalb des Studienkollektivs.

Effektmaße und klinische Relevanz

Werden am Ende der Studie zwei konkurrierende Therapiearme entlang eines kontinuierlichen primären Endpunktes ausgewertet, können die Gruppen hinsichtlich ihrer Mediane in diesem Endpunkt verglichen werden. Je größer sich dabei der Unterschied zwischen den medianen Effektmaßen der Therapiegruppen zeigt, desto höher erachtet man die *klinische Relevanz* des Studienergebnisses. Der Unterschied zwischen den Medianen misst in diesem Sinne den Effekt der neuen Therapie gegenüber der bewährten; er ist ein *Effektmaß* für den Fall eines kontinuierlichen Endpunktes. Wird stattdessen ein binärer primärer Endpunkt zu Grunde gelegt (etwa „Erhalt eines behandelten Zahnes über ein Jahr: ja / nein?"), so können statt dem Vergleich der Mediane das absolute oder das relative Risiko zwischen den beiden Therapieregimes als Effektmaß verwendet werden. In beiden Fällen werden deskriptive Maße für die Größenordnung des klinischen Unter-

[1] Bereich Klinische Epidemiologie und Gesundheitsökonomie, Rehabilitations- und Sportmedizin (Leiter: Prof. Dr. F. Krummenauer), Universitätsklinikum Carl Gustav Carus Dresden, Fetscherstr. 74, Haus 29, 01307 Dresden

[2] Klinik und Poliklinik für MKG-Chirurgie (Direktor: Prof. Dr. Dr. W. Wagner), Johannes Gutenberg-Universität Mainz, Langenbeckstr. 1, 55101 Mainz

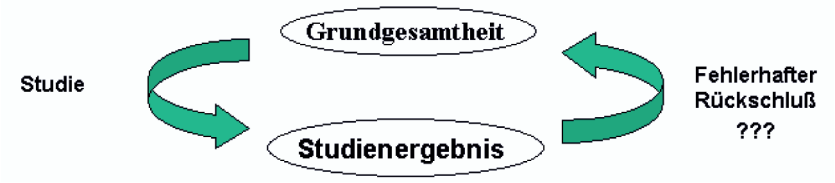

Abbildung 1 Schematische Veranschaulichung des alpha-Fehlers, der bei Übertragung von Studienergebnissen auf eine der Studie zu Grunde liegende Grundgesamtheit niemals völlig ausgeschlossen werden kann.

schiedes zwischen den beiden zu vergleichenden Regimes angegeben, welche die *klinische Relevanz* des Studienergebnisses charakterisieren.

p-Werte und statistische Signifikanz

Ein zentrales Anliegen der Biometrie besteht darin, die in einer Studie beobachteten Ergebnisse auf andere, nicht in der Studie enthaltene Patienten übertragen zu können. Da die Studie aber meist auf einer eher kleinen Auswahl (Stichprobe) von Patienten aus einem sehr viel größeren Pool (Grundgesamtheit) beruht, kann das Ergebnis der Studie anders ausfallen, als es sich tatsächlich bei den meisten anderen Patienten der Grundgesamtheit darstellt. Durch die Zufallsauswahl eines Patientenguts kann ein Studienergebnis resultieren, welches in der Grundgesamtheit gar nicht zutrifft – aufgrund der Studienergebnisse würde also möglicherweise ein falscher Rückschluss auf die Grundgesamtheit aller anderen Patienten gezogen!

Dieser potentiell fehlerhafte Rückschluss wird auch als *alpha-Fehler* oder *Fehler 1. Art* bezeichnet (Abb. 1). Klar ist, dass dieser Fehler nie gänzlich ausgeschlossen werden kann. Aber zumindest kann auf der Basis der Studienergebnisse statistisch berechnet werden, wie groß die Wahrscheinlichkeit ist, angesichts des vorliegenden Studienergebnisses diesen Fehler zu begehen. Diese statistische Wahrscheinlichkeit wird auch als *p-Wert* bezeichnet. Ist der p-Wert (%) sehr klein, so ist der alpha-Fehler „eher unwahrscheinlich" und das Studienergebnis kann statistisch gesichert auf die Grundgesamtheit aller Patienten übertragen werden. Ist er eher groß, so erscheint das Risiko eines alpha-Fehlers „eher wahrscheinlich"; das Studienergebnis kann nicht als übertragbar angesehen werden. Es sei an dieser Stelle klar betont, dass diese extrem vereinfachende Interpretation des p-Wertes in keinster Weise mathematisch valide ist, aber zumindest eine einigermaßen für Anwender zielgerichtete Arbeitsinterpretation gestattet.

In obiger Interpretation stellt sich nun die Frage, ab wann der p-Wert „hinreichend klein" ist, um eine statistisch gesicherte Aussage von einer Studie auf die Grundgesamtheit übertragen zu können? Hierfür hat sich als Qualitätsgrenze das *Signifikanzniveau* α etabliert: α gibt die maximal tolerable Wahrscheinlichkeit für einen alpha-Fehler an, welche vor Beginn der Studie (!) festgelegt werden muss. Wird α auf den gängigen Wert 5 % gesetzt, so wird für das Studienergebnis maximal eine Fehlerwahrscheinlichkeit von 5 % erlaubt. Ist es nach Ende der Studie mit einer größeren Irrtumswahrscheinlichkeit als tolerabel behaftet (ergibt sich also ein p-Wert $p > 5$ %), so kann das Studienergebnis nicht statistisch gesichert auf die Grundgesamtheit übertragen werden. Ist es mit einer geringeren Fehlerwahrscheinlichkeit als tolerabel behaftet – ergibt sich also der p-Wert der Studie zu weniger als 5 % – so kann das Ergebnis statistisch gesichert („signifikant") auf die Grundgesamtheit übertragen werden.

In diesem Sinne bedeutet statistische Signifikanz „statistisch gesicherte" Übertragbarkeit eines Studienergebnisses von der Studie auf andere Patienten. Folgende Punkte sind somit stark vereinfacht zu berücksichtigen:

- Der p-Wert ist, vereinfacht formuliert, die aus den Studiendaten berechnete Wahrscheinlichkeit, ein Studienergebnis „fälschlich" auf die Grundgesamtheit aller Patienten zu übertragen.
- Das Signifikanzniveau α ist, vereinfacht formuliert, die maximal tolerable Wahrscheinlichkeit, ein Studienergebnis „fälschlich" auf die Grundgesamtheit zu übertragen. Übliche Werte für α sind 5 %, 1 % und 0.1 %.
- Gilt $p < \alpha$, so kann, vereinfacht formuliert, das Studienergebnis modulo einer Rest-Irrtumswahrscheinlichkeit von α auf die Grundgesamtheit übertragen werden; man spricht dann von einem statistisch signifikanten Ergebnis.

Statistische Signifikanz versus Klinische Relevanz

Leider werden die Konzepte von klinischer Relevanz und statistischer Signifikanz nicht selten verwechselt. In der eingangs erwähnten randomisierten Studie [1] wurden 36 Patienten mit aggressiver Parodontitis einem scaling/root planing unterzogen. Adjunktiv erhielten 18 Patienten eine Verum-Therapie mittels Chlorhexidin-Chip (CHX) als Testmedikation, 18 Patienten der Kontrollgruppe wurden systemisch mit einer Amoxicillin/Metronidazol-Kombination (AB) behandelt. Nach sechs Monaten wurde als primärer Endpunkt die Taschen-Sondierungstiefe betrachtet; als ein wichtiger sekundärer Endpunkt wurde der klinische Attachmentlevel erhoben.

Dabei zeigte sich nach sechs Monaten zwischen den beiden Gruppen ein zum Niveau 5 % statistisch signifikanter Unterschied ($p < 0.001$) hinsichtlich der Reduktion der Sondierungstiefe [1]. Aber was besagt dieser statistisch signifikante Unterschied? Der p-Wert besagt weder, welche Therapiegruppe die stärkere Reduktion – mithin die höheren Effekte – aufweist, noch wie groß der Unterschied zwischen den Gruppen ist – es ist „lediglich" gesichert, dass der Unterschied mit hoher statistischer Sicherheit auf die Grundgesamtheit von Patienten mit aggressiver Parodontitis übertragen werden kann. Um diesen Unterschied jedoch klinisch bewer-

	Median	$Q_1 - Q_3$	p-Wert
CHX	- 1.25	- 1.51, - 0.68	
AB	- 1.91	- 2.17, - 1.24	
Differenz AB vs. CHX	0.66		< 0.001

Tabelle 1 Differenz der Sondierungstiefe [mm] nach sechs Monaten für Patienten mit adjunktiver adjuvanter Therapie mittels Chlorhexidin-Chip (CHX) vs. systemische Antibose (AB) durch Amoxicillin/Metronidazol bei aggressiver Parodontitis [1].

ten zu können, muss ein Effektmaß für den klinischen Unterschied zwischen den Therapien angegeben werden!

Die deskriptiven Ergebnisse in Tabelle 1 demonstrieren, dass der statistisch signifikante Unterschied klinisch zu einer Reduktion um im Median 1.25 mm nach CHX-Gabe und um im Median 1.91 mm nach AB-Gabe korrespondiert, also zu einem Unterschied von 0.66 mm in der Taschen-Sondierungstiefe zwischen den beiden Therapiemodi – jedoch *zugunsten der Kontrollgruppe*! Diese Tendenz des Ergebnisses ist aus dem p-Wert und der darauf basierten formalen Signifikanzbewertung nicht direkt erkennbar.

Obiger Gruppenunterschied, charakterisiert durch ein Effektmaß von 0.66 mm zwischen den Therapiemodi, besitzt ferner kaum eine klinische Relevanz für die Therapie von Patienten mit einer aggressiven Parodontitis: Die beschriebenen Effekte unterscheiden sich kaum zwischen den beiden Therapiemodi. Die „Mehr-Reduktion“ um 0.66 mm nach AB-Gabe gegenüber CHX-Gabe hat kaum einen wirklichen Mehr-Nutzen aus therapeutischer Perspektive. Der formal signifikante Unterschied zwischen Verum- und Kontroll-Therapie impliziert keinen klinisch relevanten Unterschied.

An diesem Beispiel wird sinnfällig, dass klinische Relevanz und statistische Signifikanz zwei völlig verschiedene Qualitäten eines Studienergebnis beschreiben: Der Unterschied zwischen den Therapiemodi ist im Endpunkt „Taschen-Sondierungstiefe sechs Monate nach Studieneintritt“ statistisch signifikant, hat aber kaum klinische Relevanz für die Prognose der Zahnerhaltung. Beide Informationen sind jedoch notwendig, um das Studienergebnis bewerten zu können. Dies motiviert ferner, dass Effektmaße und Maße zur Beschreibung der statistischen Signifikanz stets gemeinsam zur Beschreibung eines Studienergebnisses angegeben werden sollten. Tabelle 1 illustriert dies für die oben genannte Studie.

Korrespondenzadresse:

Prof. Dr. Frank Krummenauer
Bereich Klinische Epidemiologie und Gesundheitsökonomie, Rehabilitations- und Sportmedizin
Universitätsklinikum Carl Gustav Carus Dresden
Fetscherstr. 74, Haus 29
D-01307 Dresden
Tel.: 03 51 / 4 58 37 47
Fax: 03 51 / 4 58 43 44
E-Mail: Frank.Krummenauer@uniklinikum-dresden.de

Literatur

1 Kaner D, Bernimoulin JP, Hopfenmüller W, Kleber BM, Friedmann A: Controlled-delivery chlorhexidin chip versus amoxicillin/metronidazole as adjunctive antimicrobiell therapy for generalized aggressive periodontitis: a randomized controlled clinical trial. J Clin Periodontol 2007;34:880-891
2 Krummenauer F: Fortbildung Medizinische Biometrie IV. Klinische Monatsblätter Augenheilkunde 2002;219:817-820
3 Schumacher M, Schulgen G: Methodik Klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung, Heidelberg, Springer-Verlag 2007 (zweite Auflage)