Einsichten eines Wissenschaftsnarren (23)

Brüder, zur Sonne, dem p-Wert ein Ende, Brüder, zum Lichte empor!

Viele wollen die statistische Signifikanz via p-Wert neu definieren oder sogar ganz aus der Wissenschaft verbannen. Dabei leistet er meist gar nicht das, was sie ihm zuschreiben.

„Die Wissenschaft wehrt sich gegen die p-Wert-Tyrannei!" So zumindest verkündete es vor kurzem die *Financial Times*. Denn überall ist die Aufregung groß. Mehr als achthundert Forscher, darunter viele prominente Biostatistiker, haben dazu aufgerufen, sich gegen den p-Wert zu erheben. Und dies ist nur der Höhepunkt eines Aufstands, der schon im vergangenen Jahr begonnen hatte. Eine Gruppe von Wissenschaftlern forderte damals, dass wir die Schwelle für „statistische Signifikanz" ganz neu definieren sollten. Von derzeit meist 0,05 auf 0,005 – insbesondere wenn Wissenschaftler damit behaupten wollen, etwas entdeckt zu haben. Für viele Forscher und Experten ging diese Forderung allerdings nicht weit genug, sie fordern daher, statistische Signifikanz gleich ganz zu beseitigen, statt nur neu zu definieren. Wieso die Aufregung? Worum geht es überhaupt? Und ist das alles wirklich neu?

Wir erinnern uns: Im Jahr 2012 gewannen Craig Bennett und Kollegen mit einer bemerkenswerten Studie den Ig-Nobelpreis für Neurowissenschaften. Sie positionierten einen toten Lachs aus einem lokalen Supermarkt in einem Kernspintomographen. Dort zeigten sie dem Fisch Bilder von Menschen in sozialen Situationen mit einer bestimmten emotionalen Aufladung, etwa einen Streit oder einen Kuss. Der tote Lachs musste dann entscheiden, welche Gefühle die Abgebildeten wohl durchlebt haben mussten. Tatsächlich zeigte die Bildgebung mittels funktioneller Magnetresonanztomographie dabei signifikante Veränderungen in der Oxygenierung des toten Lachshirns – was auf eine Aufgaben-spezifische neuronale Verarbeitung im Fischgehirn hinwies.

Wie aber können „*Post-mortem*-neuronale Korrelate von Interspezies-Einfühlsamkeit im Lachs" erklärt werden, wie es der Titel des Artikels neurowissenschaftlich formuliert? Ganz einfach: Damit, dass sich die Auswertung auf statistische Standard-Signifikanzschwellen stützte und Mehrfachvergleiche nicht angemessen kontrollierte. Der Clou dabei war jedoch: Die Autoren zeigten in der Arbeit zudem, dass in 60 bis 70 Prozent der veröffentlichten funktionellen *Neuroimaging*-Studien ähnlich ausgewertet wurde – und stellten damit die Ergebnisse eines Großteils der kognitiven Neurowissenschaften in Frage.

Finden sich solche „toten Fische" vielleicht auch im Becken anderer Disziplinen, die ebenfalls stark auf multiple Testungen zurückgreifen? Etwa in Genexpressions- und -assoziationsstudien? In der Tat, auch die Genetik erkannte vor einigen Jahren – ganz ohne Ig-Nobelpreis –, dass sie ein Riesenproblem hatte: Ein Großteil der bis dato beschriebenen differenziell exprimierten Gene und Genassoziationen entpuppte sich als falsch-positive Befunde.

»Schwimmen „tote Fische" auch im Becken anderer Disziplinen?«

Zum Glück haben die Genetiker und funktionellen Hirnbildgeber mittlerweile ihre Lektion gelernt. Genetische oder Bildgebungs-Datensätze sind heute kaum noch ohne Post-hoc-Korrektur für multiple Vergleiche zu veröffentlichen. Außerdem werden, zumindest in der Genetik, Validierungen mit unabhängigen Datensätzen gefordert, bevor Assoziationen akzeptiert werden.

Das ist doch mal eine gute Nachricht, dass ganze Forschungsfelder vor ihrer Haustür gekehrt haben! Die schlechte ist jedoch, dass andernorts unzureichende Korrektur für Mehrfachtests, laxe Schwellenwerte für Typ-I-Fehler, geringe statistische Power sowie fehlende Validierung immer noch die Norm sind.

Mindestens so problematisch sind jedoch weithin verbreitete falsche Vorstellungen über das, was der p-Wert ist, und was das Label „Statistisch signifikant" bedeutet. So glauben viele Forscher, dass p die Wahrscheinlichkeit angibt, dass die Null-Hypothese wahr ist. Und folglich 1-p die Wahrscheinlichkeit, dass die alternative Hypothese (also ihre eigene Hypothese) richtig ist. Oder umgangssprachlich ausgedrückt: „Bei einem alpha von 5 Prozent laufe ich Gefahr, dass 5 Prozent meiner Hypothese trotz Signifikanz doch nicht richtig sind". Also eine Verwechslung mit der falsch-positiven Rate.

Ein weiteres häufiges Missverständnis ist, dass der p-Wert mit der theoretischen oder praktischen Relevanz des Befunds korrelieren würde. So wie der schwerwiegende Irrtum, dass die Nicht-Ablehnung der Null-Hypothese ($p > 0,05$) belegt, dass diese richtig wäre, also kein Effekt vorliegt. Und so weiter…

Aber was ist denn dann der p-Wert, und was kann er uns über unsere Ergebnisse sagen? Wenn wir die Analyse viele Male wiederholen würden und jedes Mal neue Daten generieren, und wenn die Null-Hypothese wirklich wahr ist, würden wir sie bei $p = 0,05$ in nur 5 Prozent der Fälle (fälschlicherweise) ablehnen. Mit anderen Worten: Der p-Wert stellt die Wahrscheinlichkeit dar, Daten so extrem wie (oder noch extremer als) diejenigen Ergebnisse zu erhalten, die gelten, wenn die Null-Hypothese wahr ist.

Aber klingen diese Definitionen nicht vereinbar mit der Interpretation des p-Werts als falsch-positive Rate? Schauen wir deshalb genauer hin: In den obigen Lehrbuch-Definitionen wird die Wahrscheinlichkeit auf die Daten bezogen. Ein Irrtum ist es, sie auf die Erklärung, das heißt auf die Hypothese anzuwenden. Außerdem wissen wir ja nicht, ob die Null wahr ist oder nicht. Und dann gibt es da noch das Problem der Wahrscheinlichkeit unserer Hypothese, die sogenannte *Base Rate*. Ebenso die statistische Power – das heißt die Wahrscheinlichkeit, einen Effekt zu erkennen, wenn es denn einen gibt. Dass *Base Rate* und Power für die Interpretation des p-Werts entscheidend sind, ist vielen Kollegen nicht bekannt. Und genau da liegt der sprichwörtliche Hase im Pfeffer!

Die Frage, die wir doch eigentlich gerne beantworten möchten, ist die folgende: Wenn wir einen „signifikanten" p-Wert nach einem gut durchgeführten Experiment erhalten ha-

Sämtliche Folgen der „Einsichten eines Wissenschaftsnarren" gibt es unter
www.laborjournal.de/rubric/narr

ben, mit welcher Wahrscheinlichkeit ist unser Ergebnis dann falsch positiv? Leider ist der p-Wert nur ein Teil der Gleichung, die wir lösen müssten, denn die falsch-positive Rate hängt weiterhin vom Typ-I-Fehler (alpha), dem Typ-II-Fehler (Power) sowie der Wahrscheinlichkeit der Hypothese ab, die wir testen. Je unwahrscheinlicher nämlich unsere Hypothese und je niedriger die statistische Power sind, desto wahrscheinlicher ist es, dass wir ein falsch-positives Ergebnis vor uns haben. Trotz eines signifikanten p-Werts.

Zur Verdeutlichung: Bei einem Typ-I-Fehler-Niveau von 0,05, einer Power von achtzig Prozent und einer zehnprozentigen Wahrscheinlichkeit, dass die alternative Hypothese wahr ist (also zehn Prozent *Base Rate*), sind fast vierzig Prozent der statistisch signifikanten Ergebnisse falsch positiv! Und aufgemerkt: In vielen Bereichen der Biomedizin, insbesondere in der präklinischen Forschung, liegt die statistische Power oft weit unter achtzig Prozent, eher bei fünfzig Prozent oder darunter. Und wer sich mit explorativer Forschung in wissenschaftliches Neuland vorwagt (Tun wir das nicht alle?), muss wohl auch mit *Base Rates* unter zehn Prozent rechnen. Denn sonst wäre man doch nur unoriginieller *Mainstream*-Wissenschaftler, der beforscht, was auf der Hand liegt oder was man gar schon weiß!

Foto: BIH/Thomas Rafalzyk

Ulrich Dirnagl

leitet die Experimentelle Neurologie an der Berliner Charité und ist Gründungsdirektor des QUEST Center for Transforming Biomedical Research am Berlin Institute of Health. Für seine Kolumne schlüpft er in die Rolle eines „Wissenschaftsnarren" – um mit Lust und Laune dem Forschungsbetrieb so manche Nase zu drehen.

Die Kombination aus niedriger Power, laxem Typ-I-Fehler-Niveau (alpha = 0,05), niedriger *Base Rate* und stark ausgeprägtem Bias (durch geringe interne Validität, etwa wegen fehlender Verblindung oder Randomisierung) erklärt, warum der US-Biostatistiker John Ioannidis 2005 ungestraft und seither unwiderlegt behaupten konnte, dass die meisten veröffentlichten Forschungsergebnisse falsch sein müssen.

Aber bei alpha = 0,05 ist die Wahrscheinlichkeit, einen Idioten aus sich zu machen, viel größer als fünf Prozent. Denn der p-Wert testet nicht nur die Null-Hypothese, sondern auch alles andere im Experiment.

Das schönste Beispiel hierfür ist das extrem aufwendige OPERA-Experiment, das 2011 am CERN in Genf durchgeführt wurde. Dabei gelang eine sensationelle Entdeckung: Neutrinos bewegen sich schneller als Licht! Die *New York Times* titelte damals, dass „winzige Neutrinos die kosmische Geschwindigkeitsbeschränkung durchbrochen haben". Mehrfach wurde das Experiment wiederholt, aber das Ergebnis blieb stabil bei einem p-Wert von kleiner 0,00000001. Leider führte dieser spektakuläre Befund nicht zu einem Nobelpreis, sondern zu einer totalen Blamage für die beteiligten Wissenschaftler. Wie sich später herausstellte, war ein Kabel im *Set-up* lose und ein Messinstrument war nicht richtig kalibriert. Merke: Der p-Wert bezieht sich auf die Ergebnisse eines spezifischen Experimentes und nicht auf die Hypothese! Wie spezifisch ist eigentlich Ihr Antikörper?

Der p-Wert, und damit der ganze damit verknüpfte Teststatistik-Kosmos (*Frequentist*- oder auch *Null-Hypothesis-Significance-Testing, NHST*), führt uns also schnell auf Abwege. Der p-Wert leistet nämlich meist gar nicht das, was wir von ihm erwarten – nämlich uns zu sagen, ob wir etwa Neues entdeckt haben oder ein Effekt vorliegt. Sollten wir ihn deshalb ganz aufgeben? Einfach nicht mehr testen, wie von den 800 Kollegen gefordert?

Das hieße, das Kind mit dem Bade auszuschütten! Kürzlich erst argumentierte John Ioannidis in einem Kommentar, dass „die Signifikanz (nicht nur statistisch) sowohl für die Wissenschaft als auch für das wissenschaftsbasierte Handeln wesentlich ist, und einige Filterprozesse nützlich sind, um ein Ertrinken im Rauschen der Daten zu vermeiden". Er meint damit, dass das Aufgeben von Signifikanztests unserem Bias freien Lauf lassen würde. Jeder könnte alles behaupten, und „unwiderlegbarer Unsinn würde regieren".

Wir ertrinken doch bereits jetzt in einem Meer falsch-positiver Ergebnisse. Ohne irgendeine Schwelle für die Behauptung eines Zusammenhangs oder einer Entdeckung würde sich diese katastrophale Situation mit Sicherheit weiter verschärfen. Stattdessen sollten wir strengere Regeln für die Datenerfassung und -analyse festlegen, wozu etwa die *A-priori*-Benennung und Registrierung von Hypothesen und geplanten Analyseverfahren zählen.

Obwohl weithin üblich, reicht eine Signifikanzgrenze von fünf Prozent nicht aus, um das Vorhandensein eines Zusammenhangs oder eines Effekts zu beanspruchen. Wenn überhaupt etwas, dann zeigt ein p-Wert in dieser Region, dass die Ergebnisse „einen Blick wert sind" und womöglich weitere Untersuchungen rechtfertigen – etwa eine Validierung mit größerer Fallzahl. Das Verkünden von Entdeckungen oder Effekten, die nur auf $p < 0,05$ basieren, ist grundsätzlich falsch. Und ohne ausreichende Power ist sowieso jeder p-Wert unzuverlässig, während Effektgrößen (bei einem vorhandenen Effekt) überschätzt werden.

Keines der in der aktuellen Debatte zum p-Wert vorgebrachten Argumente und kein vorgeschlagener Ausweg sind neu. Seit Einführung seiner Grundlagen durch Ronald A. Fisher, also seit fast hundert Jahren, ist er zyklisch Gegenstand von hitzigen Debatten. Auch seine Abschaffung ist schon mehrfach gefordert worden, ebenso wie die Aufgabe von NHST – also frequentistischer Statistik zu Gunsten von alternativen Ansätzen, insbesondere Bayes'scher Statistik.

»Konzentrieren wir uns lieber auf biologisches Denken.«

Auffällig ist, dass diese Diskussionen fast ausschließlich von Statistik-Afficionados geführt werden, die ohnehin wissen, wie man den p-Wert (nicht) interpretiert. Und die mit Bayes'scher Statistik vertraut sind. Viel wichtiger wäre es aber, dass *wir*, die „normalen" Forscher, uns vom Ritual der Hypothesentestung mit $p < 0,05$ verabschieden und die Interpretation unserer Ergebnisse nicht vom p-Wert abhängig machen. Dass wir uns stattdessen auf biologisches Denken konzentrieren sowie mehr Sorgfalt auf das Design, die Analyse und die Veröffentlichung unserer Studien verwenden – und dass wir diese (prä)registrieren. Methoden und Ergebnisse sollten so transparent beschrieben werden, dass Effekte und Schlussfolgerungen unabhängig bestätigt werden können.

Die Angabe von statistischen Signifikanzen ist hyperinflationär und damit bedeutungslos geworden. Teststatistiken können unsere Argumentation leiten, aber nicht bestimmen.

Weiterführende Literatur und Links finden sich wie immer unter: http://dirnagl.com/lj.