Einsichten eines Wissenschaftsnarren (62)

Trau, schau, wem – …

Wie erkenne ich Overselling, Spin und andere Merkwürdigkeiten in wissenschaftlichen Artikeln? Der Wissenschaftsnarr präsentiert eine gar nicht so närrische Checkliste.

Ein neues Jahr liegt vor uns, und wieder dürfen wir uns bis Silvester 2024 auf fast zwei Millionen neue wissenschaftliche Artikel allein in PubMed freuen! Darunter Spektakuläres, Triviales, Bestätigendes, Widersprechendes und Widersprüchliches, Redundantes, Geschöntes, Absurdes, ... – aber natürlich auch Gefälschtes und Plagiiertes. Weniges davon wird die Biomedizin revolutionieren, einiges davon wird sie voranbringen; das meiste aber wird gar nicht gelesen, geschweige denn zitiert werden. Egal ob gelesen, nützlich oder relevant – die meisten Titel werden Lebensläufe zieren und den Autoren damit zu Titeln und Fördergeldern verhelfen. Was ja eine der vornehmsten Funktionen des akademischen Publikationswesens ist.

»Welchen Publikationen kann ich trauen, bei welchen sollte ich besonders skeptisch sein?«

Wir Wissenschaftler (und die Verlage, die davon leben) haben uns in eine auf Artikeln und deren assoziierten Metriken basierte akademische Reputationsökonomie eingemauert. In ihr buhlen wir um Sichtbarkeit und Impact-Punkte. Dies produziert unweigerlich eine Paperflut. Um noch Aufmerksamkeit zu erzeugen, müssen die von uns berichteten Effekte immer größer und die behauptete Relevanz unserer Befunde für die Wissenschaft oder gar Menschheit immer wichtiger werden.

Wie navigiert man in diesem Meer von Publikationen? Gibt es eventuell sogar formale Kriterien, die uns dabei den Weg leiten könnten? Gerade jüngere Wissenschaftler – voller Ideale und noch motiviert von der Freude am Erkenntnisgewinn, dazu kaum verdorben durch die Jagd auf karrierefördernde Publikationslisten, Impact-Faktoren und h-Indizes – stellen sich diese Fragen. Ebenso wie manchmal der Wissenschaftsnarr. Was ist demnach real? Wo regiert überwiegend Spin? Welchen Publikationen beziehungsweise welchen Schlussfolgerungen darin kann ich trauen, bei welchen sollte ich besonders skeptisch sein?

Deshalb hier der Versuch einer närrischen 14-Punkte-Checkliste. Sie ersetzt weder das intensive Studium des einzelnen Artikels noch technische und inhaltliche Kompetenz. Vielleicht hilft sie aber bei einer ersten, ganz allgemeinen Einordnung des Gelesenen.

1) Außergewöhnliche Behauptungen benötigen außergewöhnliche Evidenz! Allzu häufig lesen wir über geradezu unglaubliche Entdeckungen: Orale Transplantation von Fäkalien, die im Tierversuch experimentell induzierte Schlaganfälle kurieren; Appetitsteigerung durch Handystrahlen (der Narr berichtete in *LJ* 5/22: 30-32); mediterrane Diäten, die das kardiovaskuläre Risiko stark senken, … – und so weiter und so fort. Sie wissen, welche Aussagen ich meine (Quellen und Zitate wie immer unter *http://dirnagl.com/lj*). Hier sollten wir uns immer und sogleich an Carl Sagans berühmten Spruch erinnern, der zuerst von Pierre-Simon Laplace vor über zweihundert Jahren so formuliert wurde: „Das Gewicht der Evidenz für eine außergewöhnliche Behauptung muss proportional zu ihrer Merkwürdigkeit sein."

Womit wir schon beim wichtigsten Kriterium für die Glaubwürdigkeit eines wissenschaftlichen Artikels wären: Die Qualität der darin geschilderten Evidenz. Dazu die nächsten Punkte …

2) Teststatistik und Signifikanzen: Dass ein Ergebnis „statistisch signifikant" ist, gilt in vielen Studien als das wichtigste Argument. Schon im Abstract schreit uns oftmals der p-Wert entgegen, häufig nicht einmal begleitet von der dazugehörigen Varianz oder Effektstärke. Auf letztere, und noch mehr auf deren mögliche biologische Bedeutung, kommt es aber an. Ganz zu schweigen davon, dass die ach so wichtige „statistische Signifikanz" auf tönernen theoretischen Füßen steht. Der „Erfinder" der universellen 5-Prozent-Schwelle für das falsch positive Resultat – das heißt: des Typ-I-Fehlers –, Ronald A. Fisher, formulierte die Konsequenzen bei Unterschreiten dieser Schwelle vor hundert Jahren so: „Da lohnt es sich hinzuschauen!" Mitnichten war also gemeint, eine Entdeckung zu reklamieren, und Publikationen, Doktorarbeiten oder Förderanträge darauf aufzubauen. Außerdem sollten Sie – auch wenn die Autoren das in der Regel tun –, den p-Wert nicht mit dem positiven prädiktiven Wert verwechseln. Selbst wenn die meisten Wissenschaftler das glauben, sagt der p-Wert nämlich nicht, wie wahrscheinlich der

Foto: BIH/Thomas Rafalzyk

Ulrich Dirnagl

ist experimenteller Neurologe an der Berliner Charité und ist Gründungsdirektor des QUEST Center for Responsible 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.

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

Befund ist, der damit belegt werden soll (mehr dazu in *LJ* 10/19: 24-25).

Finden sich im Artikel auch keine Aussagen zum Typ-II-Fehler – das heißt: zur statistischen Power – und keine ordentliche *A-priori*-Fallzahlabschätzung, sollten Sie doppelt skeptisch werden. Die insbesondere in präklinischen Studien sehr geringen Fallzahlen führen nicht nur zu hohen Falsch-negativ- und Falsch-positiv-Raten, sondern auch zu einer substantiellen Überschätzung der Effektstärken, falls diese überhaupt real sein sollten (sogenannter „Winner's Curse").

Und weil wir schon bei der statistischen Power sind: Wegen der in Phase 2 einer klinischen Studie recht niedrigen Fallzahlen sind diese nicht auf Wirksamkeit gepowert. Dafür macht man bei Erfolg danach eine viel größere Phase-3-Studie. Nichtsdestotrotz können viele Kliniker der Versuchung nicht widerstehen, im Rausche der statistischen Signifikanz, mit der sie ihren primären Endpunkt erreicht haben, gleich auch noch Wirksamkeit zu reklamieren. Die aber war auf diese Weise nur ein explorativer Endpunkt. Auch dies ein Warnsignal!

3) Präsentation der Daten mit Standardfehler des Mittelwerts (SEM) und Balkengraphen? Dieses Negativkriterium ist allgegenwärtig. Ich muss Sie warnen: Man will Ihnen was vormachen. Der Standardfehler des Mittelwerts wird benutzt, um eine große Streuung (Varianz) der Daten zu verschleiern; der Balkengraph – noch dazu wenn Achsen skaliert und/oder unterbrochen werden – dient der Verheimlichung der Verteilung der Daten sowie der Vorspiegelung eines substantiellen Effektes. Kampagnen wie #BarBarPlots und die Aufforderungen in den Autoreninstruktionen der Journale, echte Varianzmasse wie Standardabweichungen (SD) oder besser Konfidenzintervalle (CI) zu verwenden, sind bisher ohne Effekt geblieben. Bestehen Sie trotzdem auf Dot-Blots, Violin-Blots, Box oder Whisker-Blots sowie SDs oder CIs.

»In der „normalen" frequentistischen Statistik gibt es das Ursache-Wirkung-Prinzip nicht!«

4) Korrelation ist nicht gleich Kausalität! Manchmal direkt heraus, häufig aber subliminal verkaufen uns viele Autoren signifikante Korrelationskoeffizienten als Belege für Ursache-Wirkung-Beziehungen. A geht rauf, B geht rauf – also bewirkt der Parameter A den Parameter B. So einfach ist das aber nicht. In der „normalen" – also allgegenwärtigen – frequentistischen Statistik gibt es das Ursache-Wirkung-Prinzip gar nicht! Korrelationskoeffizienten funktionieren in beide Richtungen. A kann B bewirken, genauso wie B umgekehrt A – ein und derselbe Korrelationskoeffizient. Außerdem wirkt vielleicht C, das wir gar nicht kennen oder berücksichtigen, auf A und auf B – und stellt damit einen Scheinzusammenhang her.

Eine wunderbare Zusammenstellung solcher scheinbaren Korrelationen findet sich unter *www.tylervigen.com/spurious-correlations*. Meine Lieblingskorrelation dort ist die zwischen dem Alter der Miss America und den Morden durch Dampf und heiße Objekte. Ohne Verwendung der erst in den letzten beiden Jahrzehnten entwickelten Methoden der kausalen Inferenz mit ihren graphischen Ansätzen (vor allem den Directed Acyclic Graphs) muss weiter gelten: Ohne Intervention, nur aus der bloßen Beobachtung heraus, lässt sich eine Kausalbeziehung zwar vermuten, aber nicht belegen. Mehr dazu in einer der nächsten Folgen dieser Kolumne.

5) Wie gut war das Studiendesign? Sie sollten die folgenden Fragen mit „Ja" beantworten können: Wurden die Versuche und die Auswertung der Ergebnisse verblindet durchgeführt? Wurden die Versuchsobjekte (Zellkulturen, Mäuse, Menschen *et cetera*) randomisiert ausgewählt? Waren vorab Kriterien bestimmt und im Artikel angegeben worden, nach denen Ergebnisse in die Analyse ein- oder ausgeschlossen wurden? Berichten die Autoren dann auch, wie viele Versuchsobjekte demnach nicht eingeschlossen werden konnten – und warum? Gab es eine *A-priori*-Definition

eines relevanten Hauptergebnisses („Primärer Endpunkt"), auf den die Fallzahl-Planung ausgerichtet war? Wurden weitere Parameter vorbestimmt, die nur explorativ analysiert werden sollen („Sekundäre Endpunkte"), für welche die Studie aber nicht gepowert war? Haben die Autoren sich festgelegt, ob ihre Studie explorativ angelegt ist – also der Generierung von Hypothesen dienen soll – oder ob sie konfirmatorisch ist und damit eine Hypothese be- oder widerlegen will? Von dieser – leider selten gemachten – Unterscheidung hängen ganz wesentlich das Studiendesign und die statistischen Analyseverfahren ab. Und natürlich mehr noch aus der Studie zu erzielende Erkenntnisgewinn.

»Häufig werden in klinischen Studien sogenannte Surrogat-Endpunkte verwendet.«

6) Werden die Originaldaten im Artikel zur Verfügung gestellt? Gibt es also ein „Data Availability Statement", und was steht da drin? Ganz schlecht ist, wenn es gar keines gibt (Ausnahmen, besonders bei klinischen Studien, bestätigen die Regel). Nicht viel besser ist es, wenn es lapidar heißt: „Data available on reasonable request." Sehr gut dagegen ist ein direkter Link zum Download – unbedingt mal reinschauen! Und noch besser ist es, wenn diese Daten „FAIR" – also Findable, Accessible, Interoperable und Reusable geteilt werden sowie mit brauchbaren Metadaten annotiert sind.

7) Äußern sich die Autoren zu möglichen Interessenkonflikten? Gibt es also ein „Conflict of Interest Statement", und was steht da drin?

8) War das Studiendesign präregistriert? Bei interventionellen klinischen Studien ist dies eigentlich Pflicht – zumindest wenn die Ergebnisse später ordentlich publiziert werden sollen. Bei präklinischen Studien ist das leider noch die Ausnahme. Das ist sehr schade, denn eine Präregistrierung bringt einen massiven Qualitätssprung in der Bewertung der Evidenz, die in einem Artikel berichtet wird. Vor allem können wir uns als Leser dann versichern, dass Endpunkte und Analyseverfahren nicht *al gusto* von den Autoren im Studienverlauf verändert wurden, um erwünschte Ergebnisse zu erzielen. Denn leider müssen wir davon ausgehen, dass solches Cherry Picking von Daten sowie multiple statistische Analysen bis zum Erreichen signifikanter Resultate („Undisclosed Flexibility in Analysis and Reporting") wesentliche Gründe für die epidemische Nicht-Reproduzierbarkeit von Studienergebnissen und die Inflation von Effektgrößen sind. Präregistrierte präklinische Studien bekommen vom Narren deshalb von vornherein einen Vertrauensvorschuss! Das Gleiche gilt übrigens für Multicenter-Studien, in denen Ergebnisse unabhängig in verschiedenen Laboren repliziert wurden.

9) Gibt es eine kritische Diskussion der Limitationen? Reflektieren die Autoren die Ergebnisse und die Schlussfolgerungen, die sie ziehen, im Lichte möglicher Schwächen ihrer Studie? Dazu könnten geringe Fallzahlen, eingeschränkte Generalisierbarkeit (externe und Konstrukt-Validität), mögliche Verzerrungen (Bias), Probleme methodischer, instrumenteller und technischer Art, limitierter Zugang zu Daten, die Notwendigkeit der weiteren Absicherung durch unabhängige Konfirmation sowie vieles mehr zählen. Das Pro-forma-Auflisten von „Strohmann"-Limitationen, nur um diese dann mit einem Satz als „unwahrscheinlich" zu disqualifizieren, gilt nicht!

Die oben gelisteten Kriterien gelten für Grundlagen- wie auch für die klinische Forschung. Im Folgenden noch einige Warnsignale, auf die man besonders bei klinischen Studien achten sollte …

10) Argumentation mit relativer statt absoluter Risikoreduktion. Eine starke Reduktion des relativen Risikos durch eine neue Therapie lässt diese in sehr gutem Licht erscheinen. Wenn aber das absolute Risiko des Ereignisses gering ist – was häufig der Fall ist –, dann ist die relevante Risikoreduktion wesentlich geringer. Artikel, die die relative Risikoreduktion in den Vordergrund stellen, tun dies häufig, um uns einen großen Nutzen eines Medikamentes vorzugaukeln, der gar nicht existiert.

11) Unterscheidet sich der primäre Endpunkt im Artikel von demjenigen der Präregistrierung? Werden gar sekundäre Endpunkte plötzlich zu primären geadelt? Es lohnt sich deshalb immer, einen Blick ins klinische Studienregister zu werfen, um dies zu überprüfen!

12) Beruhen die Erfolgsmeldungen der Studie auf im Nachhinein definierten Subgruppen? Falls nicht *a priori* definiert und nicht präregistriert, kann die Analyse von Subgruppen zwar interessante Hypothesen generieren, sollte aber nicht zu positivem Spin in der Interpretation der Studienergebnisse führen.

13) Verwendet die Studie Surrogat-Endpunkte? Was für Patienten wirklich zählt („Patient Important Outcomes"), sind insbesondere Lebensqualität, Symptomfreiheit oder wenigstens -linderung sowie möglicherweise Lebensverlängerung (aber nicht um jeden Preis). Häufig werden in klinischen Studien aber sogenannte Surrogat-Endpunkte verwendet, wie zum Beispiel Veränderungen von Laborwerten, physikalischen Variablen oder in bildgebenden Verfahren. Meist wird dies mit Praktikabilität, objektiver Quantifizierbarkeit sowie pathophysiologischen Überlegungen begründet. Jedoch ist die Studienliteratur voll von Beispielen (einige davon unter *http://dirnagl.com/lj*), in denen Interventionen solche Surrogatmarker positiv beeinflusst haben, die Patienten aber letztendlich gar nichts davon hatten. Die Pharmaindustrie dagegen aber umso mehr. Denn bis die Unwirksamkeit auf Endpunkte gezeigt werden kann, die für Patienten wirklich relevant sind, kann sie Milliarden mit dem Verkauf von Tabletten verdienen, die lediglich Blutwerte verbessern.

»Abzuraten ist davon, die Reputation eines Journals als Qualitätskriterium zu verwenden.«

14) Wurde die Studie wegen Wirksamkeit vorzeitig beendet? Es erscheint paradox, dies als Warnsignal zu betrachten. In der Tat lässt sich jedoch sowohl theoretisch wie auch praktisch an einer Vielzahl von Beispielen belegen, dass es hierbei regelhaft zu einer deutlichen Überschätzung des Effektes, bei geringen Ereignisraten sogar zu falsch-positiven Resultaten kommt.

Sollten einige oder mehrere der hier angeführten formalen Kriterien erfüllt sein, heißt dies natürlich keineswegs automatisch, dass der Studie nicht zu trauen ist, dass man deren Schlussfolgerungen mit Vorsicht genießen sollte oder die Studie sonstwie von minderer Qualität ist. In jedem Fall sollten Sie beim Auftreten solcher „Warnsignale" aber noch genauer als üblich hinschauen.

Unbedingt abzuraten ist allerdings davon, die Reputation eines Journals als Qualitätskriterium zu verwenden. Ob eine Studie in *PLoS ONE* oder *Nature* veröffentlicht wurde, sagt allenfalls etwas darüber aus, für wie relevant oder spektakulär die Autoren und die Editoren die Ergebnisse halten. Und dass man sich damit häufig in die eine oder andere Richtung täuscht, zeigten zuletzt etwa der Fall und die multiplen Retraktionen des Wissenschaftsstars und nunmehrigen Ex-Stanford-Präsidenten Marc Tessier-Lavigne (siehe *LJ* 10/23: 22-24) und demgegenüber der Nobelpreis für Katalin Karikó. Ersterer veröffentlichte seine Ergebnisse vorwiegend in *Nature*, *Cell* und *Science*, Letztere im *Journal of Biochemistry*, *Molecular Therapy* und *Nucleic Acids Research*.

Bleiben Sie also skeptisch! Denn organisiertes Misstrauen produziert vertrauenswürdige Wissenschaft.

Zitierte sowie weiterführende Literatur und Links finden sich wie immer unter:
http://dirnagl.com/lj.