Die Qualität wissenschaftlicher Arbeiten – eine Bewertungshilfe für Journalisten

Gerd Antes

Für den Wissenschaftsjournalismus ist das Erkennen von fragwürdigen Forschungsergebnissen in der stetig steigenden Flut wissenschaftlicher Arbeiten eine große Herausforderung. Das Qualitätsspektrum erstreckt sich – neben den hochwertigen Arbeiten – von grundsätzlich ungeeignetem Vorgehen über kleine Schlampereien bis hin zur Fehlinterpretation von Ergebnissen und schlimmstenfalls massiven vorsätzlichen Fälschungen.

Neben den üblichen Regeln der journalistischen Recherche ist ein Grundverständnis des Wissenschaftsprozesses und der häufigsten Fehlerquellen hilfreich, wenn nicht unverzichtbar. Einen sicheren Schutz vor Fehlern kann es für den Journalisten nicht geben. Die Kenntnis häufiger Fehlerquellen und die kritische Prüfung des Quellenmaterials sind jedoch Grundlage für eine zumindest einigermaßen zuverlässige Trennung von Spreu und Weizen. Dieses Kapitel bietet dafür Unterstützung.

›Gute wissenschaftliche Praxis‹ umfasst für mich drei wesentliche Bereiche: Die ›Qualitätssicherung‹ – eigentlich eine Selbstverständlichkeit – ist real ein Problem, da der Leistungs- und Publikationsdruck in der Wissenschaft heute so hoch ist, dass die Zeit für konsequente Qualitätskontrolle fehlt. Die in der Denkschrift der *DFG* geforderte ›Ehrlichkeit und Redlichkeit‹ in der Wissenschaft kann nur erfolgreich umgesetzt werden, wenn die Inhalte aktiv dem wissenschaftlichen Nachwuchs vermittelt werden. Die ›Verantwortung des/der Wissenschaftlers/in‹ für mögliche soziale Konsequenzen und potenzielle negative Nutzung seiner Forschungsergebnisse ist durch den schnellen Transfer der Ergebnisse gestiegen, und der/die Wissenschaftler/in muss, wo nötig, auch zum ›whistleblower‹ werden.

Prof. Dr. Ulrike Beisiegel
Direktorin des *Instituts für Molekulare Zellbiologie* im *Zentrum für Experimentelle Medizin*
am Fachbereich Medizin des *Universitätsklinikums Hamburg-Eppendorf*,
Sprecherin des Ombudsman der *Deutschen Forschungsgemeinschaft* und
Mitglied des *Wissenschaftsrates*

Was ist eigentlich Wissenschaft?

Sucht man nach der Definition von Wissenschaft, so findet man zahlreiche unterschiedliche Formulierungen. Verzichtet man auf die Diskussion verschiedener Wissenschaftsansätze und ihrer philosophischen Grundlagen, gelangt man zu einer pragmatischen, für dieses Kapitel ausreichenden Charakterisierung.

Hauptziel der Wissenschaft ist die rationale, nachvollziehbare Erkenntnis der Strukturen, Zusammenhänge, Abläufe und Gesetzmäßigkeiten der natürlichen wie der historischen und kulturell geschaffenen Wirklichkeit. Wissenschaft beruht besonders auf dem Prozess methodisch betriebener Forschung und der Darstellung der Ergebnisse sowie der angewendeten Methoden. Daraus resultiert »ein System des durch Forschung und Lehre und überlieferte Literatur gebildeten, geordneten und begründeten, für sicher erachteten Wissens einer Zeit« (Brockhaus 2003: 2193). Der Bezug auf eine bestimmte Zeit ist Ausdruck der Vergänglichkeit des Wissens, das in immer kürzeren Zyklen durch wissenschaftliche Aktivitäten umgewälzt und erneuert wird.

Wesentliches Element wissenschaftlicher Arbeit ist daher der Diskurs. Ziel muss es immer sein, bestehendes Wissen (inhaltlicher wie methodischer Art) zu bestätigen (Verifikation) oder zu widerlegen (Falsifikation) und durch neues Wissen zu ersetzen. Wissenschaft ist also ein kumulativer Prozess. Ein Prüfstein für die Beurteilung neuen Wissens ist deswegen der Bezug zum bereits vorhandenen Wissen. Systematische Übersichtsarbeiten, die den kumulierten Wissensstand aus verschiedenen Arbeiten zusammenfassen und gewichten sollen, werden seit über 20 Jahren als methodisches Werkzeug entwickelt und angewendet (Chalmers et al. 2002; Egger et al. 2001; Khan et al. 2004).

Wissenschaft wird oft als objektiv und wertfrei bezeichnet. Diesem Anspruch mögen auch heute noch einige wenige Forscher nahekommen, die mit ausreichend privatem finanziellem Polster in weitestmöglicher Unabhängigkeit und mit hohem Ethos offene Fragen wissenschaftlich zu beantworten suchen. Die Realität sieht jedoch für die überwiegende Mehrheit – vor allem in den Natur- und Ingenieurwissenschaften und speziell auch in der Medizin – völlig anders aus.

Infrastrukturelle Voraussetzungen (und damit Finanzmittel) wie Laborfläche oder Rechnerleistung, die als Steuerungsinstrument bewusst eingesetzte leistungsorientierte Mittelvergabe sowie der globale Wettbewerb sind Faktoren, denen der Wissenschaftsbetrieb heute fast überall ausgesetzt ist. Forschungsfinanzierung, soziale Strukturen der Forscherlandschaft und weitere Faktoren üben massiven Einfluss auf die Auswahl und die Durchführung von wissenschaftlichen Arbeiten aus.

Objektivität ist somit für sich genommen kein geeigneter Begriff für die Bewertung von wissenschaftlicher Arbeit, da er unmittelbar in eine komplexe Diskussion führt, was darunter überhaupt zu verstehen ist. Stattdessen hat die Forderung nach Transparenz in den letzten Jahren eine zentrale Rolle eingenommen, um die Integrität von Forschungsarbeiten sicherzustellen.

Die Betrachtung der Rahmenbedingungen für eine Forschungsarbeit sollte Bestandteil jeder journalistischen Recherche sein, da damit ein Indikator für die erste vorsichtige Einschätzung der Glaubwürdigkeit von Forschungsergebnissen gegeben ist.

Allgemeine Qualitätsmerkmale wissenschaftlicher Publikationen: ›Peer review‹ und ›Impact factors‹

Das ›Peer review‹-Verfahren, also die Begutachtung einer wissenschaftlichen Arbeit durch Gleichgestellte (Peers) vor der Entscheidung über eine Veröffentlichung, gilt allgemein als das Qualitätsmerkmal wissenschaftlicher Zeitschriften. Dabei wird meistens übersehen, dass ›Peer review‹ ein äußerst heterogener Vorgang ist und keineswegs automatisch ein fehlerfreies Produkt liefert.

Empirische Untersuchungen des ›Peer review‹ werden alle drei Jahre auf einer internationalen Konferenz mit fast allen Herausgebern namhafter Zeitschriften präsentiert (*Fifth International Congress on Peer review and Biomedical Publications 2005*). Die Fachwelt ist sich über die Defizite im Klaren (siehe die Beiträge von Gerhard Fröhlich und Holger Wormer in diesem Buch), da es aber bisher keine erfolgversprechenden Alternativen des ›Peer review‹ gibt, kann man es als das am wenigsten schlechte Verfahren für die Bewertung im Wissenschafts- und Publikationsprozess ansehen. Es ist auch notwendig als pragmatisches Steuerungsinstrument für die Auswahl jener nur sieben bis acht Prozent aus den eingesandten Manuskripten, die in einer Zeitschrift wie beispielsweise dem *British Medical Journal (BMJ)* schon aus Platzgründen nur publiziert werden können. Die Kenntnis der Mängel bedeutet für den Leser (und somit auch für den Journalisten), dass die Verantwortung nicht abgegeben werden kann und die Qualität einer Publikation auch in hochrangigen Zeitschriften immer wieder aus eigener Perspektive bewertet werden muss.

Voraussetzung dafür ist weitgehende Transparenz, sodass in diesem Bereich die Transparenzforderung eine enorme Bedeutung erlangt hat. Als einen Schritt zu größerer Transparenz sehen viele Forschungsorganisationen und Wissenschaftler das System des ›open access‹ an, da damit der vollständige freie Zugang zur publizierten Wissenschaft ermöglicht wird (Berliner Erklärung 2003). Ob das die geeignete Antwort auf die Krise der wissenschaftlichen Literatur sein wird, ist offen; bemerkenswert ist jedoch, dass auch bei ›open access‹-Zeitschriften meist ein klassischer ›Peer review‹ durchgeführt und als Qualitätsmerkmal angesehen wird, wie z. B. bei *PloS* (www.plos.org), *BiomedCentral* (www.biomedcentral.com).

Der zweite, die Wissenschaftswelt dominierende allgemeine Qualitätsmaßstab ist der *Science Citation Index (SCI)* der privaten Firma *Thomson Scientific* (www.thomson.com/solutions/scientific/), dem Nachfolger des *Institute for Scientific Information (ISI)*. Dieses Maß für die durchschnittliche Zitierhäufigkeit von Artikeln in einer Fachzeitschrift (in den folgenden zwei Jahren nach ihrer Veröffentlichung) hat einen ungeheuren Einfluss bekommen, da es als Steuerungsinstrument für einen zunehmend höheren Anteil der Ressourcenverteilung in Fakultäten und Forschungsinstitutionen dient – und das, obwohl man auch miserable Artikel in einer Zeitschrift mit hohem ›Impact factor‹ finden kann. Somit ist man hier ebenfalls nicht sicher vor ernsthaften Fehlern in Artikeln dieser Zeitschriften und sollte wiederum seiner eigenen Urteilskraft trauen.

›Peer review‹, ›Impact factors‹ und des Weiteren die Aufnahme in Literaturdatenbanken wie die *Medline* in den *Life Sciences* (www.pubmed.gov) sind heute dominante Faktoren für die Bewertung von wissenschaftlichen Artikeln. Doch obwohl sie als Qualitätsmerkmale gelten, bieten sie nur wenig Schutz, selbst gegen schwerste Fehler in Artikeln aus hoch bewerteten Zeit-

Für mich besteht die Aufgabe der Wissenschaft in erster Linie darin, Neues zu entdecken. Wissenschaftliche Forschung soll Zusammenhänge aufdecken, die bisher unverstanden waren, und Werkzeuge schaffen, mit denen wir bewältigen können, was zuvor unmöglich war. Naturwissenschaftliche Grundlagenforschung schafft das ›Saatgut‹ für die angewandte Forschung und Ingenieurskunst von morgen. Wir müssen die Wissenschaft fördern, um unsere Lebensqualität und Wettbewerbsfähigkeit in der Zukunft zu sichern. Wissenschaft braucht den Dialog mit der Öffentlichkeit, um junge Menschen für diese schöne Aufgabe zu begeistern und um die notwendigen Geldmittel zu mobilisieren.
Prof. Theodor W. Hänsch
Physiker am *Max-Planck-Institut für Quantenoptik* in Garching bei München und der *Ludwig-Maximilians-Universität*, Physiknobelpreisträger 2005

schriften. Es lohnt sich daher, auch über diese eher formalen Faktoren hinaus ein gewisses Rüstzeug für die Bewertung wissenschaftlicher Arbeiten zu erwerben. Für einzelne Fachgebiete müssen jeweils spezielle Datenbanken (wie z. B. *Chemical Abstracts* für die Chemie) beachtet werden, während andere Datenbanken (wie z. B. Biosis) naturwissenschaftliche Grundlagenartikel enthalten oder wie der *Science Citation Index* allgemein interdisziplinär ausgerichtet sind.

Unvermeidlich: systematische Fehler und Variabilität in empirischen Forschungsarbeiten

Jede journalistische Recherche sollte – im Idealfall – eine möglichst fehlerfreie Erkundung und entsprechend eindeutige Darstellung eines Themas anstreben. Dieser journalistische Wunsch nach Klarheit und Schutz vor Fehlern bedeutet aber ein Dilemma, denn er stößt in der Wissenschaft auf eine Welt, die sich gerade gegensätzlich verhält: Wissenschaft fordert geradezu zu Fehlern auf, indem bestehendes Wissen infrage gestellt sowie neues Wissen gesucht wird und dabei vorsätzlich Unsicherheit und Irrwege in Kauf genommen werden.

Tabelle 1: Vermutete Wirkungen

Wirksam – oder was?	
• Klinisch getestet	Hilft gegen …
• Bewährt/akzeptiert/etabliert	Hat positiven Einfluss auf …
• Lange Tradition	Dermatologisch bestätigt
• Anerkannte Behandlungsmethode (traditionelle chinesische Medizin)	Wirkt auf Nieren und Leber
• Liefert vielversprechende Ergebnisse	Regeneriert die Zellen
• Macht stressfest; gibt uns Kraft	… sind Heilmittel gegen Stress
• Geben dem Körper die Kraft, sich zu reinigen	baut die Schleimhäute auf (Magen)
• Entgiftet den Körper	Viel Vitamin C, E, Kalium, Eisen
	Können Krebs hemmen
	Hält Blutgefäße geschmeidig
	Stärkt das Immunsystem

Zusätzlich erschwerend kommt hinzu, dass vermeintliches Wissen sich vielfach nur auf allenfalls pseudowissenschaftliche Begründungen stützt, die allerdings oft umso vehementer wiederholt werden, je vager sie sind. Selbst purer Glaube wird regelmäßig durch Aussagen wie »lange bewährt« oder »durch wissenschaftliche Studien belegt« mit der Aura von Wissenschaftlichkeit umhüllt (siehe Tabelle 1 mit einer Sammlung solcher Begriffe).

Die Überprüfung solcher Aussagen kann schwierig sein und erheblichen Aufwand bedeuten, schon weil die ›Studien‹ selten konkret benannt sind. Oft wird dieser Aufwand jedoch nicht betrieben, sondern die vermeintlich belegten Aussagen werden kritiklos weitergegeben. Wie weit verbreitet das ist, zeigt schon das folgende (harmlose) Beispiel:

Beispiel 1: Konservierung einer halb geleerten Flasche Sekt

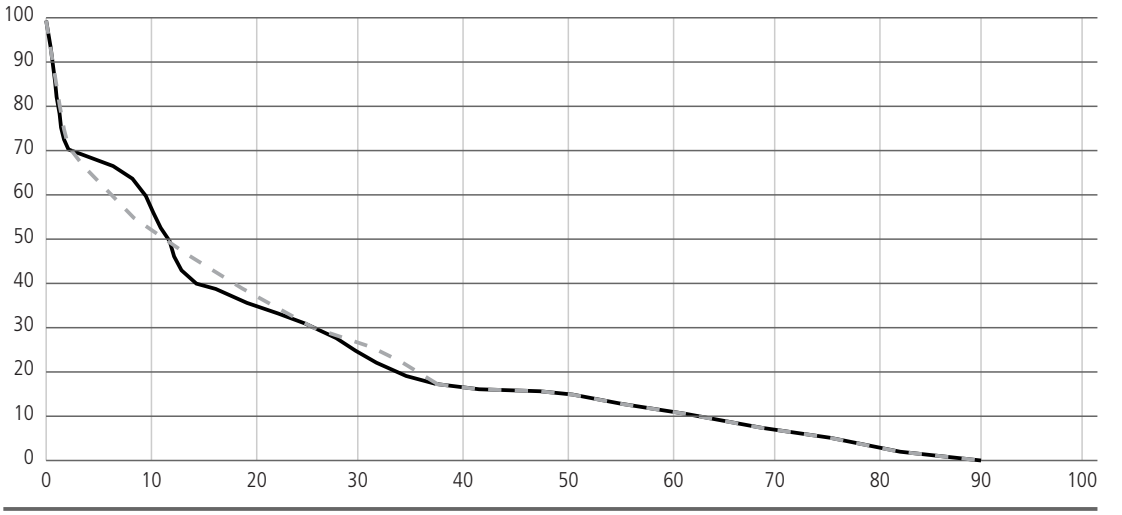
Möchte man eine halb geleerte Flasche Sekt am nächsten Abend in gleicher Qualität, also mit gleichem Kohlensäuregehalt bzw. gleicher Perligkeit, weitertrinken, so besteht ein weit verbreitetes Hilfsmittel darin, die Flasche mit einem umgekehrt in den Flaschenhals gesteckten Teelöffel (nach Möglichkeit aus Silber!) offen in den Kühlschrank zu stellen. Dieses Verfahren ist schätzungsweise 80 bis 90 Prozent der deutschen Bevölkerung bekannt und von einem hohen Anteil (vermutlich über 70 Prozent) schon einmal angewendet worden, wie der Autor jedenfalls in einer Vielzahl von spielerisch angelegten Befragungen völlig unterschiedlicher Auditorien in den letzten Jahren ohne Ausnahme feststellen konnte.

Die Überprüfung des Nutzens dieser Maßnahme ist einfach und verkörpert die Prinzipien moderner empirischer Forschung: Eine gewisse Anzahl Sektflaschen wird halb geleert, die Hälfte davon mit einem Löffel ›verschlossen‹, und alle werden unter gleichen Bedingungen im Kühlschrank aufbewahrt, bis nach einem Tag ein möglicher Unterschied geprüft wird. Für diesen Test wird der Sekt mit neutralem Etikett abgefüllt und verkostet und seine Perligkeit auf einer geeigneten Skala bewertet. Dieses Experiment wurde vor einigen Jahren durchgeführt und erbrachte das eigentlich nicht verblüffende Ergebnis, dass die Anwendung des Löffels keinerlei messbaren Effekt hat, sondern die Abnahme der Kohlensäure mit oder ohne Löffel gleich schnell erfolgt (Abbildung 1).

Dieses scheinbar triviale Beispiel ist ein weitreichendes Modell dafür, wie unter viel ernsthafteren Bedingungen die kritiklose Anwendung von tradiertem Wissen – im günstigsten Fall – nicht den erhofften Nutzen bringt, darüber hinaus aber Schaden verursachen kann. Doch auch ohne direkten Schaden besteht immer ein (gerne übersehener) negativer Effekt darin, dass nutzlos verbrauchte Ressourcen andere nützliche Maßnahmen verhindern.

Gerade in der medizinischen Versorgung wird das zunehmend erkannt und hat zu rigorosen Forderungen nach Wirksamkeitsnachweisen durch klinische Studien geführt. Solche Wirksamkeitsstudien wie auch viele andere Untersuchungen sind in ihrer Struktur der Prüfung des ›Löffel-Effekts‹ sehr ähnlich, sodass dieses Beispiel immer parat sein sollte, wenn man Zweifel an einem behaupteten Effekt hat.

Abbildung 1: Einfluss eines Löffels für die Konservierung von Sekt in einer geöffneten Flasche



Angaben in Stunden

Quelle: New Scientist, May 2000

Bias und Zufall

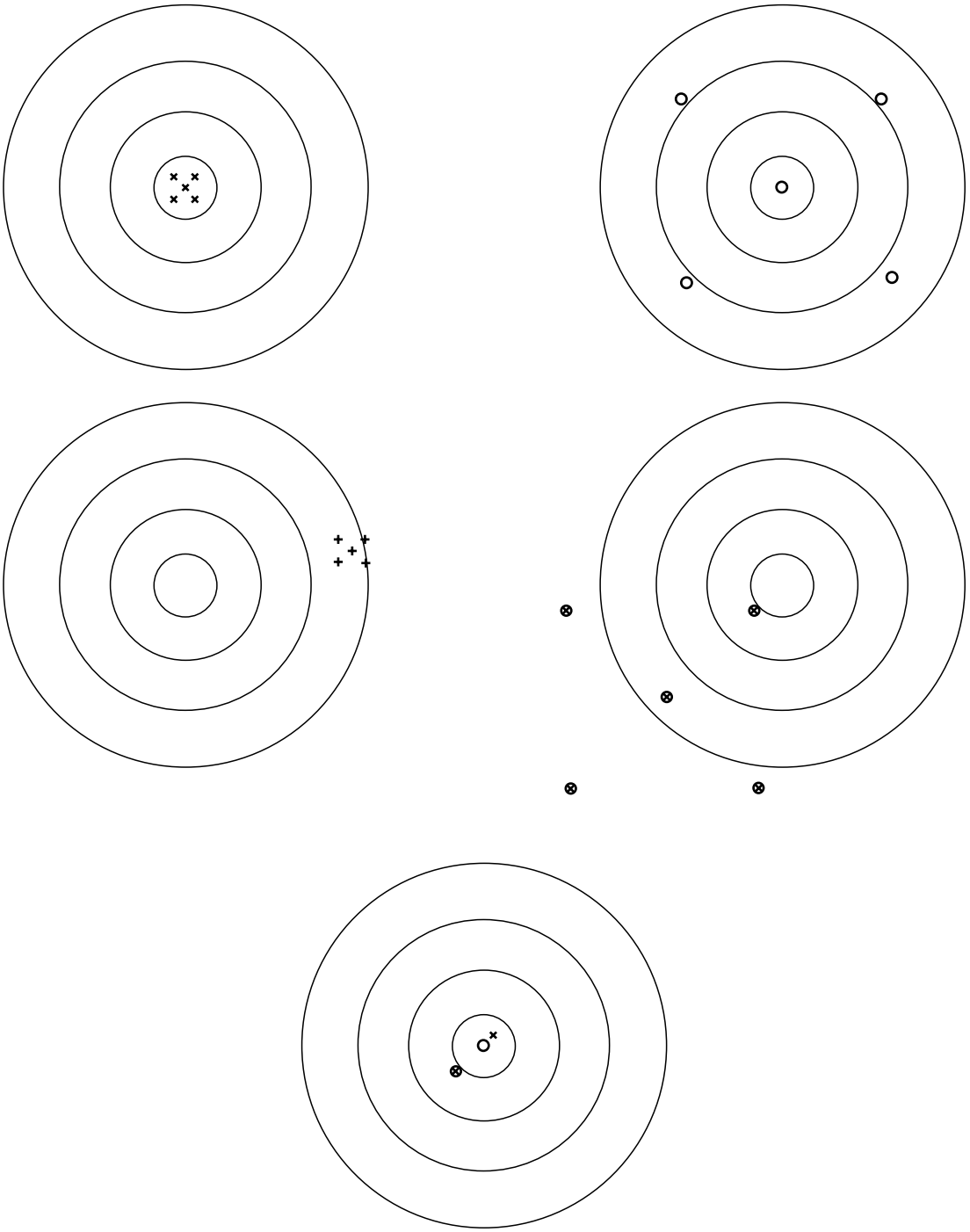
Das Leben in der wissenschaftlichen Welt der Unsicherheit fordert einen sorgfältigen Umgang mit ihr. In der empirischen Forschung sind dafür zwei Begriffe von zentraler Bedeutung:

Der ›Bias‹ ist ein Sammelbegriff für systematische Fehler oder Verzerrungen, der aus der englischen Sprache stammt und heute in vielen Diskussionen benutzt wird, allerdings oft fehlerhaft. Eine weitere, in ihrer Natur grundsätzliche andere Fehlerursache rührt daher, dass jedes Experiment und jede Beobachtung zufälligen Schwankungen (›play of chance‹) unterliegt und damit auch unter gleichen Bedingungen bei Wiederholung nicht zu gleichen Ergebnissen führt (›zufälliger Fehler‹/›Streuung‹). Das Zusammenspiel dieser beiden Fehlerquellen lässt sich an einem einfachen Beispiel verdeutlichen.

Beispiel 2: Schießen auf eine Zielscheibe

In Abbildung 2 ist das Ergebnis von vier Serien aus jeweils fünf Schüssen auf eine Zielscheibe in schematisierter Form dargestellt. Schütze 1 zeigt weder einen systematischen noch einen großen zufälligen Fehler. Schütze 2 unterliegt starken Zufallsfehlern, ist jedoch auch im Mittel treffsicher. Schütze 3 verfehlt das Ziel mit hoher Präzision, während Schütze 4 den gleichen systematischen Fehler mit dazu noch starken Zufallsschwankungen zeigt.

Abbildung 2: Schematische Darstellung des Einflusses von Bias und Streuung



Quelle: Eigene Darstellung

Dieses Beispiel kann als Modell für Experimente oder Erkenntnisprozesse dienen und zeigt die Schwierigkeiten bei der Identifikation von Fehlerursachen. Besonders folgenreich kann die Situation mit Schütze 3 sein, da die hohe Präzision dazu verleiten kann, den systematischen Fehler, also einen Bias, nicht zu erkennen. In dem Beispiel könnte das ein Fehler des Schießgeräts wie auch ein systematischer, reproduzierbarer Fehler des Schützen sein.

In der Realität sind Unterschiede nicht so leicht zu erkennen wie in dieser idealisierten Darstellung, da die ›Wahrheit‹ (oft auch als Goldstandard bezeichnet), die in dem Beispiel durch die konzentrischen Ringe der Zielscheibe beschrieben wird, nicht bekannt ist. Die Beobachtung der fünften Scheibe veranschaulicht, wie irreführend einzelne Beobachtungen sein können. Nur durch die Markierungen der Treffer ist erkennbar, dass die Treffer unter völlig unterschiedlichen Bedingungen erzielt wurden. Ohne diese Markierungen könnte man von einer homogenen Situation ausgehen und daraus sehr fehlerhafte Schlüsse ziehen. Zur Abgrenzung der vier Situationen gegeneinander ist eine ausreichend große Anzahl Wiederholungen notwendig. Dadurch werden Punktwolken produziert, durch deren Lage und Ausdehnung systematische und zufällige Einflüsse besser erkannt und unterschieden werden können.

Bias – und was Schutzmaßnahmen wie ›Verblindung‹ dagegen bewirken können

Bias kann in vielfältiger Form auftreten. Seine Kontrolle und weitestmögliche Reduzierung ist eine der zentralen Herausforderungen bei der Planung, Durchführung und Auswertung wissenschaftlicher Studien. Für das Verständnis dieses erst einmal abstrakt erscheinenden Konzepts ist ein Blick auf die wissenschaftliche Untersuchung der Wirksamkeit eines Arzneimittels hilfreich.

In einer klinischen Studie zum Vergleich zweier Medikamente werden die Probanden in zwei Gruppen eingeteilt und jeweils entweder mit Präparat A (neu) oder B (alt) behandelt. Wählt der Arzt nun jeweils die schwerer erkrankten Patienten für die Behandlung mit A aus, da er das neue Medikament für wirksamer hält, würde ein solcher Vergleich mit einem schweren Bias behaftet sein und Medikament A benachteiligen, da es den Heilungserfolg unter erschwerten Bedingungen erzielen muss. Ein solcher Selektionsbias ist eine häufige und oft sehr schädliche Fehlerursache, die Ergebnisse wertlos machen kann. Diese Studie hätte bereits in ihrer Anlage einen schweren Bias.

Bias muss also bereits im Design vermieden werden. In diesem Fall ist der geeignete Schutzmechanismus die zufällige Zuweisung der einzelnen Patienten zu einer Behandlung (Randomisierung). Nicht der Arzt, sondern ein Zufallsmechanismus wählt die Behandlung aus. In diesem Fall spricht man von einer ›randomisierten kontrollierten Studie‹, die heute aufgrund ihrer hohen Immunität gegenüber dem Selektionsbias eine der am häufigsten angewendeten Studienformen ist. Ob die Behandlungsgruppen ausbalanciert sind bezüglich Einflussfaktoren, wie z. B. dem Alter, wird in der modernen Literatur in der sogenannten ›Table 1‹ dargestellt. Damit wird auf einen Blick ersichtlich, ob z. B. die Altersverteilung in den Gruppen nahe beieinanderliegende Mittelwerte hat und damit eine häufige Verzerrungsquelle ausgeschaltet ist. Das Vorhandensein dieser Tafel sollte als Teil des Qualitätschecks immer geprüft werden.

Eine weitere, nicht nur in klinischen Studien lauernde Bias-Ursache stammt von den oft starken Erwartungen der Beteiligten, und zwar sowohl bei den Ärzten wie auch den Patienten. Die geeignete Schutzmaßnahme dagegen ist die möglichst weitgehende ›Verblindung‹ der Beteiligten, die im einfachsten Fall der Arzneimittelstudien dadurch erreicht wird, dass die Medikamente beider Gruppen neutral verpackt werden und nur anhand einer Nummer identifiziert werden können, deren Dekodierung bis Studien-Ende in einem Safe ruht. Ärzte wie Patienten sind in Unkenntnis der tatsächlichen Therapie; deswegen wird die Studie als ›doppelblind‹ bezeichnet.

Diese Form der Bias-Abwehr ist für die Untersuchung anderer Therapieformen (z. B. für Operationstechniken) nur teilweise möglich. Oft werden in solchen Fällen nur die Patienten unwissend gehalten (›einfachblind‹), während auf der ärztlichen Seite Behandlung und Beurteilung getrennt von zwei Ärzten erfolgen, damit der Arzt nicht seinen eigenen Behandlungserfolg beurteilt.

Auch die Methode der Verblindung kann man sich in Analogie zum Eingangsbeispiel mit dem Silberlöffel klarmachen: Wer an die Methode glaubt, wird beim Verkosten der Flaschen, die mit Löffel ›konserviert‹ wurden, womöglich dazu neigen, bei ihnen eine höhere Perligkeit zu schmecken, ohne dass dies tatsächlich der Fall ist. Auch hier müsste man den Verkoster deswegen für den Test ›verblinden‹.

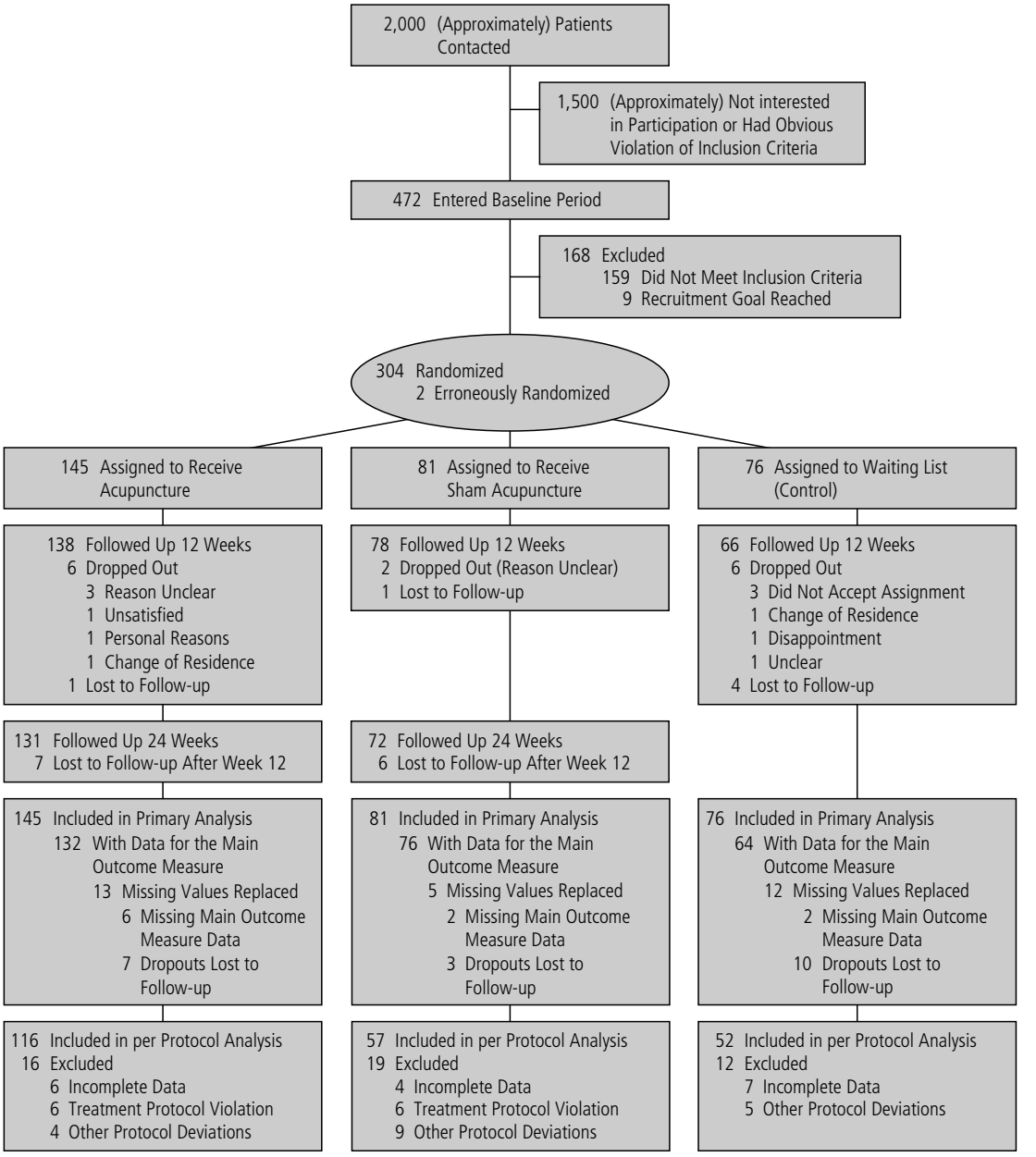
Eine der folgenträchtigsten Bias-Ursachen ist die fehlerhafte Analyse des vorzeitigen Ausscheidens aus einer Studie. Patienten brechen aus diversen Gründen ihre Studienteilnahme ab, sodass mit bis zu 30 Prozent Verlust gerechnet werden muss. Der Umgang mit diesen Abbrechern bietet die Möglichkeit, die Ergebnisse einer Studie zu verzerren und – falls es vorsätzlich geschieht – nach Belieben zu manipulieren. Aus diesem Grund muss ein Diagramm mit lückenloser Beschreibung des Studienverlaufs aller eingeschlossenen Patienten unverzichtbarer Bestandteil eines jeden Studienreports sein. Dieses Diagramm (in jüngeren Publikationen oft die Figure 1) ist ein zentraler Qualitätsparameter, der in jeder Veröffentlichung einer Studie gesucht werden sollte (siehe Abbildung 3).

Die Kontrolle solcher Bias-Ursachen hat eine so extreme Bedeutung, dass sie in jeder wissenschaftlichen Arbeit im Methodenteil explizit behandelt werden sollte. Für die journalistische Arbeit bedeutet das, den Blick auf jeden Fall auf die Methodenbeschreibung zu werfen. Auch wenn die technischen Details nicht voll verstanden werden, erlaubt der Eindruck von diesem Abschnitt eine Einschätzung, ob die Arbeit grundsätzlich vertrauenswürdig erscheint. Eine Hilfe bei der Überprüfung ist das sogenannte ›consort statement‹ (siehe unten).

Streuung – und wie man dem Einfluss des Zufalls gerecht wird

Eine sich grundsätzlich vom Bias unterscheidende Fehlerursache rührt von vielfältigen zufälligen Einflüssen her. Experimente, Beobachtungen oder Studienergebnisse erhalten dadurch eine Komponente, die bei einer möglichen Wiederholung zwangsläufig anders ausfällt. Stellt man sich vor, die Studie würde mehrfach wiederholt, so würden die Studienergebnisse um das wahre Ergebnis streuen. Numerisch wird die Streuung durch den sogenannten Standardfehler oder die Varianz beschrieben. Die Streuung fällt umso größer aus, je kleiner die Studie

Abbildung 3: Flussdiagramm des Studienverlaufs für die Gesamtheit der zu Beginn aufgenommenen Patienten in einer Studie zur Wirksamkeit von Akupunktur bei Migräne: drei Behandlungsgruppen mit ungleicher Besetzung (doppelte Stärke in der Akupunktur-Gruppe)



* Although two patients assigned to the acupuncture group came only to the initial examination but did not return at the end of the baseline phase, one of the large study centers erroneously registered them for randomization. According to the analysis plan the intent-to-treat population comprised only patients with baseline data.

Quelle: Linde 2005

ist. Eine große Streuung bedeutet eine geringe Präzision der Studienergebnisse. Deswegen müssen Studien ausreichend groß sein, damit die Studienergebnisse eine genügend präzise Schätzung der (unbekannten) Wahrheit erlauben.

Die erreichte Präzision wird dargestellt, indem die aus der Studie ermittelten Zahlenwerte mit Grenzen versehen werden, die sich mit statistischen Methoden errechnen lassen. Das dadurch definierte Intervall heißt Konfidenzintervall (C.I.), da es ein Intervall angibt, in dem man den wahren Wert mit einem bestimmten Vertrauen erwarten kann. Das Vertrauen wird üblicherweise so bestimmt, dass der wahre Wert in dem zwischen den Grenzen liegenden Intervall mit 95-prozentiger Wahrscheinlichkeit erwartet werden kann.

Ein Ergebnis von 3,8 (C.I. 3,2–4,4) bedeutet also, dass der geschätzte Wert von 3,8 einigermaßen sicher (nämlich nur mit fünf Prozent Wahrscheinlichkeit) nicht außerhalb des Intervalls 3,2–4,4 liegt (Fletcher 2007). Die 95-prozentige Wahrscheinlichkeit ist eine reine Konvention, mit der das Vertrauen in Aussagen quantifiziert wird. Diese Zahl ist die Kehrseite der ebenfalls konventionellen Festlegung von fünf Prozent für die Fehlerrate von Signifikanztests. Das heißt, ein Signifikanztest auf dem Niveau von fünf Prozent lässt bei 100 Wiederholungen im Mittel in fünf Fällen auf die Überlegenheit eines Medikaments schließen, auch wenn dies nicht der Wahrheit entspricht.

Konfidenzintervalle sind Signifikanztests (p-Werte) in ihrer Berechnung äquivalent, aber vorzuziehen, da Tests nur Ja-Nein-Entscheidungen erlauben, während die Konfidenzintervalle quantitative Aussagen z. B. zu einem Therapieeffekt liefern. Ein Konfidenzintervall kann prinzipiell für jede (!) Schätzung aus einer Studie bestimmt werden. Die Berechnung erfolgt meistens mit Standardprogrammen und ist in einfachen Fällen problemlos mit einem Taschenrechner möglich.

Konfidenzintervalle und die vor Beginn einer Studie notwendige Berechnung der Studiengröße sind essenzielle Qualitätsparameter, die bei Betrachten einer Publikation zur ersten Einschätzung sofort gesucht werden sollten. Als Faustregel sei an dieser Stelle bezüglich der Studiengröße nur erwähnt: Je kleiner beispielsweise der zu erwartende Wirkungsunterschied zwischen zwei Medikamenten ist, desto größer muss eine Studie sein, um den Unterschied mit ausreichender Sicherheit zu finden. Diese Sicherheit bezieht sich auf die statistische Aussagesicherheit, die jedoch sorgfältig von der klinischen Relevanz zu unterscheiden ist.

Durch eine sehr große Studie kann selbst ein äußerst kleiner Unterschied, wie z. B. wenige mm Hg bei der Blutdruckmessung, statistisch belegt werden. Das als Überlegenheit eines Medikaments zu interpretieren wäre jedoch völlig unangemessen, wenn der Unterschied klinisch völlig unbedeutend ist. Inhaltliche medizinische Aspekte sollten über den oben beschriebenen methodischen Anforderungen auf keinen Fall vergessen werden!

Wie sich bereits anhand der Darstellung von Aspekten wie ›Bias‹ und ›Streuung‹ zeigt, enthalten klinische Studien also viele methodische Einzelheiten, die vom recherchierenden Journalisten nicht schnell durchschaut werden können. Das Abprüfen der Existenz bestimmter Schritte in Planung und Durchführung und die Beschreibung davon in der Publikation ist dennoch ein wirksames Mittel der Einschätzung. Leitlinien für die Autoren von Studienreports, wie z. B. dem *CONSORT Statement for Reporting Randomized Trials* (www.consort-statement.org), bieten dem Leser eine hervorragende Hilfe für die Überprüfung, ob der Report die wesentlichen Punkte enthält.

Ähnliche standardisierte Instrumente entstehen gegenwärtig für andere Studientypen wie diagnostische Studien, durch die die Treffsicherheit sowie die Fehlerraten von diagnostischen Verfahren bestimmt werden, oder auch für Beobachtungsstudien, die Daten nur durch Beobachtung erheben und nicht durch aktive Intervention in Abläufe eingreifen. Auch sie werden als Checklisten für die schnelle Überprüfung von Studienberichten dienen und damit ein wesentliches Instrument für die Qualitätsbewertung sein. Inzwischen gibt es auch für andere Studientypen, wie z. B. Beobachtungsstudien, Qualitätsvorgaben für die Studienberichte, die in jeweils ähnlicher Weise durch speziell angepasste Checklisten als Hilfe für eine schnelle Qualitätsprüfung von Artikeln dienen können. Als Einstiegsseite zu diesen verschiedenen Berichtsleitlinien wird www.equator-network.org empfohlen.

Signifikanztests, p-Werte, Fall-Kontrollstudien, Metaanalysen und viele weitere Begriffe aus dem Methodenarsenal wirken sicherlich entmutigend auf den, der sich um eine Qualitätsbewertung einer Arbeit bemüht. Für eine Einschätzung der komplexen Darstellung im Methodikteil einer Publikation ist die Beratung durch einen mit dem Forschungsgegenstand nicht direkt verbundenen Methodiker zu empfehlen. Hilfreich für die Begriffserläuterungen sind Glossare wie das internationale Standardwerk von John M. Last (2001) oder frei zugänglich und deutschsprachig das Glossar des *Deutschen Netzwerks Evidenzbasierte Medizin* (EbM-Glossar des *EbM-Netzwerks*). Als nicht technische, mit vielen Beispielen arbeitende Bücher können für allgemeine epidemiologische Zusammenhänge Robert H. Fletcher et al. (2007) und für klinische Studien Martin Schumacher und Gabi Schulgen (2007) empfohlen werden.

Vom Ergebnis her betrachtet, sind es vor allem drei Dinge, die gute Wissenschaft auszeichnen: fundamental neue Erkenntnisse, lesenswerte Veröffentlichungen und nicht zuletzt hervorragend qualifizierter wissenschaftlicher Nachwuchs. Die Grenzen gesicherten Wissens zu überschreiten, neue Methoden zu entwickeln und ein bislang unbekanntes Forschungsterritorium zu erkunden erfordert Fantasie, Risikobereitschaft und Standvermögen. Für eine Förderinstitution sind zudem durch Expertenrat unterstützte Auswahlverfahren, großzügige und verlässliche Finanzierung des Besten sowie Vertrauen in die Geförderten und schließlich Geduld mit Blick auf das Erzielen bahnbrechender Ergebnisse unerlässlich.
Dr. Wilhelm Krull
Generalsekretär der *Volkswagen Stiftung* in Hannover

Häufige Fehler in der Präsentation und Interpretation wissenschaftlicher Ergebnisse

In der wissenschaftlichen Welt wie auch in deren Darstellung in den Medien tauchen einige typische Fehler besonders häufig auf und können Ursache für schwerwiegende Irreführungen sein, obwohl sie eigentlich als leicht durchschaubar erscheinen. Im Folgenden werden eine Reihe der häufigsten Fehler sowohl von wissenschaftlicher wie auch von journalistischer Seite besprochen und – soweit es sie gibt – Schutzmaßnahmen dagegen empfohlen.

Vorsicht vor logischen Kopfständen:
Aussagen und Empfehlungen sollten den Grundregeln der Aussagenlogik folgen

Wissenschaftliche Aussagen sollten immer den Grundsätzen der formalen Aussagenlogik folgen, die eigentlich leicht einsehbar sind, jedoch oft missachtet werden. So kann man Allaussagen nicht beweisen, sondern nur widerlegen. Um etwa eine Aussage wie »Jeder Kupferdraht leitet Strom« zu beweisen, müsste man alle Kupferdrähte auf der Welt prüfen. Ein einziger nicht leitender Draht genügt jedoch bereits, um die Aussage als falsch zu erkennen. Andererseits kann man Existenzaussagen nur beweisen bzw. bestätigen, jedoch nicht widerlegen. Für den Beweis der Aussage »Es gibt das Ungeheuer von Loch Ness« genügt das Vorzeigen des Ungeheuers. Für die Widerlegung seiner Existenz müsste man den See leer pumpen, was nur theoretisch möglich ist. Aussagen über die Nichtexistenz kann man nicht beweisen. Eine kurzweilige Lektüre zu dieser Thematik bietet Christoph Bördlein (2002).

Dieses Regelwerk wird strikt in den theoretischen Wissenschaften, vor allem natürlich in der Mathematik, angewendet, bildet jedoch auch die Grundlagen für die Methodik der empirischen Forschung. Bei der Arbeit mit Daten wird die Situation sehr viel komplizierter, da dann statistische Aussagen nach diesen Regeln verknüpft werden müssen.

Einige medizinische Beispiele mögen dies verdeutlichen:

Eine Tollwutinfektion gilt als sicher tödlich, wenn nicht innerhalb kurzer Zeit nach der Infektion eine Impfung erfolgt. Diese Allaussage müsste demnach durch eine einzige Heilung widerlegt werden können. In der realen Welt sind hier jedoch sofort Einschränkungen zu machen. Ist »gilt als sicher tödlich« wirklich eine Allaussage oder doch vielleicht nur mit 99-prozentiger Sicherheit wahr? Damit wäre die Bedrohung für einen Infizierten zwar fast genauso groß, eine beobachtete Heilung wäre jedoch keinesfalls eine Sensation, die die These von der Unheilbarkeit zu Fall bringt, sondern nur etwas, was es auch vorher schon gab (wenn auch selten). Die zweite Fehlerquelle ist, dass die Diagnose nicht richtig war, also gar keine Tollwutinfektion vorkam.

Ähnlich liegt der Fall bei den immer wieder behaupteten Wunderheilungen. Hier ist es weit verbreitet, durch Einzelfallbeschreibungen von Heilungen (also Existenzaussagen) den Eindruck zu erwecken, dass der Heilungsmechanismus allgemein wirkt und nicht nur in dem speziellen Fall. Die einzig richtige Beschreibung kann daher nur über die Heilungswahrscheinlichkeit erfolgen, die nur durch ausreichend große klinische Studien ermittelt werden kann. Wie im vorherigen Beispiel wird man der vermeintlichen ›all-or-nothing‹-Situation mit deterministischen Aussagen nicht gerecht, sondern muss die Fehler bzw. Unterschiede durch eine geeignete statistische Betrachtung mitberücksichtigen.

Das aber gilt auch für den umgekehrten Fall, denn besonders häufig sind Fehlinterpretationen, wenn es ›keinen‹ Nachweis für den Nutzen eines Therapieverfahrens gibt. Denn damit ist keinesfalls der fehlende Nutzen nachgewiesen, sondern es ist nur der misslungene Versuch des Nutzennachweises festzuhalten. Im modernen Fachjargon heißt das: »Die fehlende Evidenz für einen Therapieeffekt ist nicht gleich der Evidenz für einen fehlenden Therapieeffekt!«

Also: Vorsicht vor scheinbar plausiblen Aussagen! Die Prüfung auf den logischen Gehalt erlaubt meistens schnell Aufschlüsse über die grundsätzliche Zuverlässigkeit einer Aussage.

Die Welt ist nicht monokausal: Vorsicht ›confounder‹!

Einfache Erklärungen sind besser zu verkaufen als komplexe. Fehlerhafte Vereinfachungen sind deswegen eine permanente Bedrohung. Sehr häufig werden Phänomene in einen monokausalen Zusammenhang hineingezwängt, der der Realität nicht entspricht. Scheinbare Zusammenhänge beruhen darauf, dass dahinterliegende bekannte oder unbekannte Ursachen – sogenannte ›confounder‹ (Störfaktoren) – die wahre Ursache sind: Apfelproduzenten können ihr Produkt mit gesundheitsfördernden Leistungen schmücken, wenn man nicht berücksichtigt, dass Personen, die viele Äpfel essen, womöglich auch sonst gesünder leben und mehr Sport treiben; die Schuhgröße kann fälschlicherweise das Einkommen bestimmen, wenn man übersieht, dass das Geschlecht meist die Schuhgröße und das Einkommen bestimmt; der kürzlich berichtete protektive Effekt von Hausarbeit auf Brustkrebs ist vermutlich eher die Folge von unbekannten sozioökonomischen Hintergrundvariablen als ein direkter Zusammenhang (Fux 2006).

Mit statistischen Verfahren lassen sich diese Zusammenhänge ausgleichen, allerdings nur, wenn sie bekannt sind. Der große Aufwand für randomisierte kontrollierte Studien hat seinen Grund genau darin, den Einfluss von ›confoundern‹ zu minimieren. Einfache Beobachtungsstudien sind durch den verzerrenden Einfluss von ›confoundern‹ erheblich mehr gefährdet.

Also: Vorsicht vor dem Tunnelblick auf nur eine Einflussgröße, da die Welt hochdimensional und komplex ist. Abfragen, ob ›confounder‹-Einflüsse bei der Wahl des Studiendesigns und in der Analyse berücksichtigt wurden!

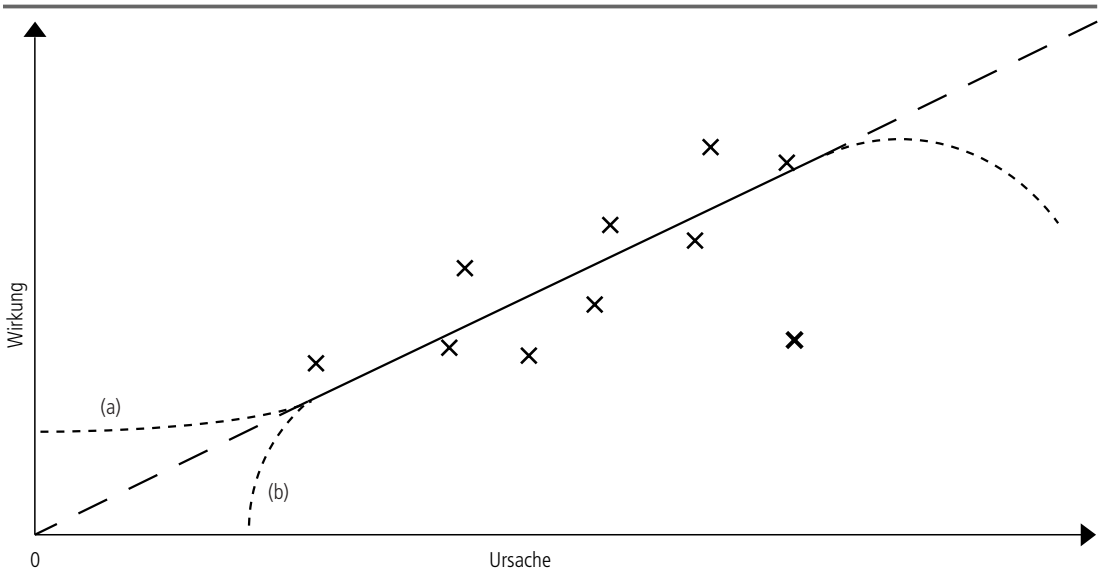
Störche et al.: Gleichzeitiges Auftreten bedeutet nicht kausale Ursache

Die Suche nach kausalen Begründungen ist eine der zentralen Triebfedern der Forschung. Entsprechend groß ist die Versuchung, das beobachtete (zufällig) zeitgleiche Auftreten zweier Phänomene in eine Ursache-Wirkung-Beziehung umzudeuten. Ein viel zitiertes Beispiel beweist, dass kleine Kinder vom Klapperstorch gebracht werden: Die Geburtenrate ist gesunken, und die Anzahl Störche hat deutlich abgenommen, also, so die Folgerung, werden die Kinder vom Storch gebracht.

Obwohl scheinbar banal und damit leicht vermeidbar, wird diese Schlussweise in offener oder verkappter Form an unzähligen Stellen benutzt, um beim Leser eines Artikels (sowohl in Fachzeitschriften wie in Publikumsmedien) den Eindruck eines kausalen Zusammenhangs zu hinterlassen. Dieses Beispiel lässt sich auch im vorigen Abschnitt als ›confounder‹-Problem formulieren. Zivilisatorische Einflüsse sind Ursache für beide Entwicklungen (bei den Störchen wie bei den Kindern) – und damit ist letztlich die Zeit der ›confounder‹.

Also: Korrelationen (oft auch als Scheinkorrelationen bezeichnet) nicht als Kausalzusammenhang interpretieren!

Abbildung 4: Unterschiedliche Modelle für die Extrapolation in Bereichen ohne gemessene Daten



Quelle: Eigene Darstellung

Kühne Schlüsse durch Extrapolation in Regionen ohne Daten

Experimente und Beobachtungsstudien liefern Daten in bestimmten Bereichen, für die dann ein Zusammenhang zwischen verschiedenen Größen durch eine Kurve dargestellt wird (Abbildung 4). Oft betrifft die relevante Fragestellung jedoch nicht die gemessene Region, sondern Randbereiche, für die keine Daten vorhanden sind. Ein typisches Beispiel ist der Einfluss von Partikeln in der Luft auf Erkrankungen der Atemwege, wofür der Niedrigdosisbereich entscheidend ist.

Gute Wissenschaft soll aus Neugierde und zum Wohle der Menschheit und der Natur ausgeführt werden. Das ist aber nicht so einfach, denn die Kenntnisse, die die Grundlagenforschung liefert, können sowohl zum Positiven als auch zum Negativen benutzt werden, und das wird wohl immer so bleiben. So gesehen gibt es keine nur ›gute Wissenschaft‹, aber wir sollten alles daransetzen, uns dem Ideal einer nur ›guten Wissenschaft‹ zu nähern.

Paul Crutzen
Nobelpreisträger für Chemie (1995) und Professor emeritus am *Max-Planck-Institut für Chemie*, Mainz

Die Kurve wie (a) beginnen zu lassen setzt die Annahme voraus, dass auch bei einer Nulldosis bereits eine Wirkung vorhanden ist. Im Gegensatz dazu wird mit (b) vorausgesetzt, dass erst ab einer Dosis überhaupt eine Schädigung eintritt. Die Frage nach der Existenz einer Schwellendosis, unterhalb der keine Gefährdung mehr gegeben ist, wird mangels Daten oft

zu einer weltanschaulichen Diskussion, die eher von ökonomischen Interessen als von Wissenschaftlichkeit bestimmt wird.

Entsprechend sieht es in der anderen Richtung aus. Aus dem schnurgeraden Anstieg des Dax kann nicht geschlossen werden, dass sich diese Entwicklung ein weiteres halbes Jahr so fortsetzt.

Solche Analysen bauen zwangsläufig auf Annahmen, die oft willkürlich sind und damit die Extrapolation spekulativ machen. Der Verzicht auf Extrapolation ist allerdings keine Option, da die Bereiche ohne gemessene Daten oft gerade enorme praktische Relevanz haben. Entscheidend ist, die zugrunde liegenden Annahmen deutlich zu machen.

Also: Vorsicht bei der Extrapolation von auf Datenbasis ermittelten Trends in Bereiche, für die keine Daten vorliegen!

Hinterher ist man immer schlauer

Für die Anlage, Durchführung und Diskussion von Studien sind die zeitlichen Zusammenhänge von großer Bedeutung. Prospektive und retrospektive Studien unterscheiden sich darin, welche Position vom Beobachter jeweils eingenommen wird. Vor allem in Konflikten werden die zeitlichen Zusammenhänge bei der Wissensgenerierung oft nicht ausreichend berücksichtigt (»Das hätte man doch schon früher berücksichtigen müssen«).

Wissenschaft ist ein kumulativer Prozess, der von einem sich weiterentwickelnden Stand der Kenntnis geprägt wird. Vollmundige Erklärungen, dass man etwas immer schon gewusst habe, entbehren oft jeder wissenschaftlichen Grundlage. Eher ist der Fall, dass ein vormals vorhandener Glaube heute durch Daten belegt ist. Ein Beleg für die prognostische Fähigkeit des Glaubenden ist es jedoch keinesfalls.

Also: Immer den zeitlichen Standpunkt prüfen und die jeweilige Perspektive betonen.

Zufallsbefunde haben nur selten Bedeutung: die Versuchung von Untergruppen

Auch große, gut angelegte Studien zeigen oft nicht die gewünschten Ergebnisse. Der Druck, aufregende Ergebnisse zu produzieren und diese dann in einer hochrangigen Zeitschrift zu publizieren, führt Forscher in die Versuchung, nach Untergruppen zu suchen, sodass zumindest dort bei einer Teilpopulation das gewünschte Ergebnis nachgewiesen werden kann. In ausreichend großen Studien mit vielen gemessenen Variablen ist es nur eine Frage der Ausdauer, bis eine signifikante Aussage in einer Untergruppe gefunden wird.

Subgruppenanalysen sind legitim, wenn sie a priori geplant und im Studienprotokoll festgehalten sind. Sie sind akzeptabel, wenn sie nach Studienende ungeplant durchgeführt werden, dieses nur zur Hypothesengenerierung und nicht konfirmativ interpretiert wird und dies vor allem in der Publikation so dargestellt wird. Es ist jedoch unredlich, nur die Subgruppenanalyse darzustellen und Umfeld und Gesamtstudie zu unterschlagen.

Also: Vorsicht vor Ergebnissen, die aus Teilen der Studienpopulation gewonnen wurden! Falls solche Ergebnisse präsentiert werden, müssen sie entsprechend gekennzeichnet werden.

Zwischenauswertung: Der Sieger wird an der Ziellinie ermittelt

In gleicher Weise gefährlich wie die Subgruppenanalysen ist die wiederholte, vielleicht sogar regelmäßige Betrachtung der Ergebnisse einer Studie während ihrer Durchführung. Die während des Verlaufs zufällige Entwicklung der Ergebnisse erlaubt bei regelmäßiger Betrachtung, gewünschte Ergebnisse zu einem geeigneten Zeitpunkt auszuwählen – so als würde man ein Zielfoto nicht erst nach einer festgelegten Distanz schießen, sondern dann, wenn der gewünschte Kandidat vorne liegt. Durch eine solche Zwischenauswertung oder Interimsanalyse werden alle in die Planung eingeflossenen wahrscheinlichkeitstheoretischen Überlegungen wertlos gemacht.

Auch hier gilt, dass eine a priori geplante Interimsanalyse mit entsprechend angepassten statistischen Eigenschaften legitim und oft sogar wünschenswert ist.

Also: Vorsicht vor ungeplanten Zwischenauswertungen! Entscheidend für die Auswertungszeitpunkte ist die Planung, wie sie im Studienprotokoll festgelegt ist.

»Garantiert nebenwirkungsfrei« bedeutet »garantiert wirkungsfrei«

Die Entwicklung eines Medikaments bedeutet eine Gratwanderung, um einerseits den Nutzen zu maximieren und andererseits das potenzielle Risiko von Nebenwirkungen so klein wie möglich zu halten. Bei wirksamen Arzneimitteln und anderen Therapien können unerwünschte Nebenwirkungen nie (!) ausgeschlossen werden. Deswegen ist es umso erstaunlicher, wie häufig Verfahren als »garantiert nebenwirkungsfrei« beschrieben werden. Träfe das zu, würde es auch »garantiert wirkungsfrei« bedeuten, da erwünschte und unerwünschte Wirkungen nicht prinzipiell zu trennen sind, sondern nur in oft jahrelangen Prozessen in ein günstiges Verhältnis gesetzt werden können.

Der einseitige Blick auf positive Wirkungen ist ein Phänomen, das bei Medikamenten, Diagnoseverfahren und Screeningmaßnahmen regelmäßig beobachtet werden kann, jedoch mit dem Anspruch an die qualifizierte Bestimmung einer Nutzen-Schaden-Relation nicht vereinbar ist.

Also: Vorsicht vor ausschließlich positiver Beschreibung von Verfahren! Die Frage nach Nebenwirkungen ist kein Suchen nach dem Haar in der Suppe, sondern die notwendige Berücksichtigung der Realität.

Prozent- oder Absolutzahlen – was auch immer eindrucksvoller ist

Auch bei völlig identischer Situation lässt sich durch die unterschiedliche Darstellung der zahlenmäßigen Zusammenhänge und Ergebnisse beim Betrachter eine sehr unterschiedliche Wahrnehmung erzeugen. Marginale absolute Unterschiede werden in prozentualer Darstellung eindrucksvoll aufgebläht: Wenn jemand beispielsweise von einer »Halbierung des Risikos« spricht, diese Änderung jedoch von 2 : 100.000 zu 1 : 100.000 erfolgt, so weist die Prozentangabe allein auf eine Bedeutung hin, die nicht gegeben ist. Da 2 : 10 und 1 : 10 die

gleiche Risikorelation bedeutet, aber wesentlich relevanter wäre, ist die Angabe von absoluten Zahlen unverzichtbar, um die tatsächlichen Risikoverhältnisse zu verstehen.

Andererseits werden oft nur Absolutzahlen angeben, um Gefahren zu betonen. 8.000 an einer Krebsart Neuerkrankte in Deutschland sehen bedrohlich aus, sind jedoch nur ein Zehntausendstel der deutschen Bevölkerung, sodass hier die Bezugsgröße wesentlich ist. Ebenso sinnlos sind alleinstehende Absolutzahlen bei der Angabe der Verkehrstoten in einem Bundesland, der durch Haie getöteten Badenden oder der im Haushalt Verunglückten, wenn keine Bezugsgrößen für einen Vergleich vorhanden sind.

Also: Bei der alleinigen Angabe von Prozentzahlen oder Absolutzahlen jeweils die fehlenden Angaben beschaffen, um die vollständige Beschreibung der Situation zu ermöglichen.

Fazit

Für die wissenschaftsjournalistische Recherche gibt es kein Patentrezept, mit dem man die Qualität des Recherchegegenstands zuverlässig einschätzen kann. Checklisten für die Abfrage einzelner Qualitätsparameter können eine wertvolle Hilfe sein, verlangen jedoch ein Grundverständnis der Rationale hinter den einzelnen zu prüfenden Begriffen.

Das einzige durchgängige Konzept für die Qualitätsbewertung ist die Fähigkeit, die vielfältigen Möglichkeiten eines Bias zu erkennen. Selbst in Wissenschaftskreisen wird diese Perspektive jedoch oft nicht voll verstanden und deswegen nicht angemessen berücksichtigt. Erkenntnisverfälschende Bias-Ursachen, die noch durch zufällige Fehler überlagert werden, stellen deswegen eine besondere Herausforderung für den recherchierenden Journalisten dar. Mit einem Grundverständnis von wissenschaftlichen Studien und deren Schwachstellen ist es möglich, sich in diesem Gebiet zurechtzufinden und die lauernden Fußangeln zu vermeiden.

Literatur

»Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen«. Conference on Open Access to Knowledge in the Sciences and Humanities. 20.–22.10.2003, Berlin. 22.10. 2003. www.zim.mpg.de/openaccess-berlin/BerlinDeclaration_dt.pdf (Download 10.4.2007).

Bördlein, Christoph. *Das sockenfressende Monster in der Waschmaschine. Eine Einführung ins skeptische Denken*. Aschaffenburg 2002.

Brockhaus. »Wissenschaft«. *Der Brockhaus Naturwissenschaft und Technik*. Bd. 3. Leipzig 2003. 2193.

Chalmers, Iain, Larry V. Hedges und Harris Cooper. »A brief history of research synthesis«. *Evaluation & The Health Professions* (25) 1 2002. 12–37.

»EbM-Glossar des EbM-Netzwerks«. www.ebm-netzwerk.de/grundlagen/glossar#glossar_html (Download 19.4.2007).

Egger, Matthias, George Davey-Smith und Douglas G. Altman. *Systematic Reviews in Health Care: Meta-Analysis in Context*. 2. Auflage. London 2001.

»Fifth International Congress on Peer review and Biomedical Publications«. Chicago, USA 2005. www.ama-assn.org/public/peer/peerhome.htm (Download 10.4.2007).

Fletcher, Robert H., Suzanne W. Fletcher und Edward H. Wagner. *Klinische Epidemiologie. Grundlage und Anwendungen*. Bern 2007.

Fux, Christiane. »Hausarbeit beugt Brustkrebs vor«. *Focus online* 29.12.2006. www.focus.de/gesundheit/ratgeber/krebs/news/studie_nid_41756.html (Download 19.4.2007).

Heilmann, Klaus. *Das Risiko der Sicherheit*. Stuttgart und Leipzig 2002.

Khan, Khalid, Regina Kunz, Jos Kleijnen und Gerd Antes. *Systematische Übersichtsarbeiten und Meta-Analysen. Ein Handbuch für Ärzte in Klinik und Praxis sowie Experten im Gesundheitswesen*. Berlin 2004.

Last, John M. *A Dictionary of Epidemiology*. New York 2001.

Linde, Klaus, et al. »Acupuncture for Patients With Migraine. A Randomized Controlled Trial«. *JAMA* (293) 17 2005. 2118–2125.

Schumacher, Martin, und Gabi Schulgen. *Methodik klinischer Studien*. Berlin 2007.