

»Während einer nur Zahlen und Zeichen im Kopf hat,
kann er nicht dem Kausalzusammenhang auf die Spur kommen.«
(Arthur Schopenhauer, Philosoph)

»Unmerklich beginnt man die Fakten so zu drehen,
dass sie sich den Theorien anpassen,
anstatt die Theorien den Fakten.«
(Sherlock Holmes, fiktiver Detektiv)

»Die hinterhältigste Lüge ist die Auslassung.«
(Simone de Beauvoir, Philosophin)

3 Vom Umgang mit Zahlen

Zahlen, Daten, Fakten – das klingt nach handfesten unumstößlichen Wahrheiten, nach den Tatsachen, an denen es nichts herumzudeuten gibt. Wissenschaftliche Veröffentlichungen bestehen zu einem Gutteil aus Ziffern, Zahlen, Daten. Ob in der Wirtschaft, der Physik, Soziologie oder der Medizin – Forschungsergebnisse und Risikoberechnungen, Zuwachsprognosen und Wirksamkeitsnachweise kommen gern als Zahlen daher. Angesichts von Zahlenreihen und Kurven zu erkennen, was bemerkenswert, aussagekräftig oder gar falsch ist, und den nüchternen Daten Leben einzuhauchen, gehört zum Handwerk des Wissenschaftsjournalisten.

Der Pressedienst erspart seinen Nutzern die mühselige Ziffernlektüre und berichtet von enormen Zuwächsen, geringem Risiko und überschaubaren Kosten. Die Adjektive gelten als anschaulich. Und sie liefern die Deutung gleich mit.

In der Wissenschaft ist aber Präzision gefragt. Ob die Messreihen, die ein Physiker liefert, die Daten aus klinischen Studien und Schulvergleichen oder Modelle in der Wirtschaftswissenschaft, sie bestehen hauptsächlich aus Zahlen, die in wissenschaftlichen Veröffentlichungen auch abgedruckt werden müssen, damit man die Versuche und die Schlüsse daraus nachvollziehen kann. Sie werden durch den Wissenschaftler interpretiert – so oder so.

Der Journalist, der darüber berichtet, sollte die Zahlen kennen und verstehen, um entscheiden zu können, ob die Zuwächse tatsächlich enorm, die Risiken gering und die Kosten überschaubar sind. Das klingt banal, ist es aber nicht. Zum einen braucht es ein wenig statistisches Können und rechnerisches Handwerkszeug, um Zahlen zu »lesen.« Zum anderen braucht es gesunden Menschenverstand, um mög-

liche Fallen zu erkennen. Nicht immer nämlich belegen die Daten das Ergebnis, das die Autoren aus ihnen herauslesen.

Zuletzt braucht es schreiberische Kniffe, die Essenz der Daten präzise, aber für Laien verdaulich aufzubereiten.

Will das was heißen? Die Signifikanz

Zahlen, die Schlagzeilen machen, haben in der Regel eines gemeinsam: Sie sind »statistisch signifikant«. Raucher leiden signifikant öfter an Lungenkrebs, Pflanzen, die mit Pferdemist gedüngt werden, bringen signifikant mehr Ertrag, und Kinder, die gestillt wurden, sind später nicht signifikant weniger von Allergien geplagt. Dieser viel strapazierte Begriff bedeutet Folgendes: Die Wahrscheinlichkeit, dass dieses Ergebnis purer Zufall ist und keinen tatsächlichen Zusammenhang belegt, liegt unter 5 Prozent. Andersherum liegt die Wahrscheinlichkeit dafür, dass die gemessenen Unterschiede (mehr Lungenkrebs, mehr Ertrag, genauso viel Allergie) etwas mit dem untersuchten Faktor zu tun haben, bei 95 Prozent. Sie ist nicht absolut, aber hoch genug, so dass man sich auf dieses Kriterium geeinigt hat.

Die Signifikanz zu berechnen, gehört zu den leichtesten statistischen Übungen. Auch wenn man nur sporadisch in die Verlegenheit gerät, selbst nachzurechnen, so sollte man doch wissen, wie das geht: mit dem Vier-Felder-Test.

Der Vierfeldertest

(auch Chi-Quadrat-Test, engl. Crosstab oder Fourfold Test)

Dieser einfache statistische Test ist ein wichtiges Mittel, um zu prüfen, ob Unterschiede in zwei Gruppen, die größer als 6 sind, zufällig sind. Die vier Felder sehen so aus:

	Ereignis	Nichtereignis	Summe
Stichprobe 1	a	b	a + b
Stichprobe 2	c	d	c + d
Summe	a + c	b + d	n

$$\text{Prüfgröße} = \frac{n (a \cdot b - b \cdot c)^2}{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}$$

Statistisch signifikant sind die Unterschiede zwischen den beiden Gruppen, wenn die Wahrscheinlichkeit, dass die Unterschiede rein zufällig auftreten, kleiner ist als 5 Prozent. Für eine Irrtumswahrscheinlichkeit von $p < 0,05$ liegt der kritische Prüfwert bei 3,841. Im Klartext: Wenn die Prüfgröße kleiner als 3,841 ist, so liegt hier kein statistisch signifikantes Ergebnis vor.

Ein typisches Beispiel ist ein Vierfeldertest für ein neues Medikament. Hier nehmen 60 Patienten an einer klinischen Studie teil. Die Hälfte bekommt das neue Mittel, die andere die Standardtherapie. Nach Ablauf der Testphase sind 11 Patienten durch die Standardtherapie geheilt worden; das neue Medikament kurierte 19. Um zu testen, ob das Ergebnis dieser Studie statistisch signifikant ist, setzen wir die Zahlen in die vier Felder ein und berechnen die Summen.

	geheilt	nicht geheilt	Summe
Standardtherapie	11	19	30
neues Medikament	19	11	30
Summe	30	30	60

Jetzt errechnet man die Prüfgröße nach der Formel oben:

$$\begin{aligned}
 \text{Prüfgröße} &= \frac{60 \cdot (11 \cdot 11 - 19 \cdot 19)^2}{30 \cdot 30 \cdot 30 \cdot 30} \\
 &= \frac{60 \cdot 57.600}{810.000} \\
 &= \frac{3.456.000}{810.000} \\
 &= 4,3
 \end{aligned}$$

Mit 4,3 ist die Prüfgröße größer als 3,841. Daraus folgt: Der Test ist statistisch signifikant. Auf Deutsch: Das neue Medikament heilt statistisch signifikant mehr Menschen und ist damit besser als das herkömmliche.

Doch auch wenn die Signifikanz noch so sauber berechnet wurde, heißt das: Bei 5 von 100 Studien liegt kein echtes Ergebnis vor, sondern die Daten lassen es zufällig so erscheinen, als hätte z. B. eine neuartige Krebsbehandlung die erwünschte Wirkung.

Als wäre das nicht Besorgnis erregend genug, untersucht kaum eine Studie nur einen einzigen Parameter. Ein Parameter ist das Alter, das Gewicht, Blutzucker, Lebensqualität, Zellmengen und Ähnliches, was sich gut messen lässt. Wenn zwei unabhängige Parameter untersucht werden, so gilt für jeden einzelnen die 5-Prozenthürde zur statistischen Signifikanz. Damit steigt das Risiko auf 10 Prozent, dass einer ein falsches (zufälliges) Ergebnis zeigt.

Ein beliebtes Beispiel für statistischen Nonsens sind Rechenübungen mit Störchen und Geburten. Eine Untersuchung aus Berlin zeigt einen signifikanten Zusammenhang zwischen Storchpopulation und Geburtenrate (Höfer 2004): Schrumpft die Storchpopulation in einer Region, so gehen dort auch die Geburtenzahlen zurück, stagniert die Zahl der Störche, stagnieren die Geburten. Sollte sich ein solcher signifikanter Zufall nicht gleich einstellen, so lassen sich die Geburtenraten mit den Populationen von 10 oder 20 Vogelarten vergleichen, und wenn das nicht reicht, ein paar Amphibienarten. Die Wahrscheinlichkeit wird schon für irgendeine Zahlenreihe einen Zusammenhang zeigen. Wenn nicht mit dem Storch, so hängen die Geburtenzahlen rein statistisch möglicherweise an der Geburtshelferkröte.

Bevor Sie allzu sehr auf wissenschaftlich Bewiesenes vertrauen, sollten Sie wissen, dass mindestens 5 Prozent der Studien keinen echten Effekt nachweisen. Wenn man in einer umfangreichen Studie 20 falsche Hypothesen testet, so wird eine davon als richtig herauskommen.

Prozente, Prozentel

Rabattaktionen, Entwicklungen, Verbesserungen, Chancen und Risiken werden in Prozenten angegeben. Prozentzahlen klingen oft sehr aussagekräftig und dramatisch und sind doch in den allermeisten Fällen irreführend.

Ein fiktives Beispiel:

Schaufensterkrankheit ist der umgangssprachliche Name einer Durchblutungsstörung, die Menschen zwingt, wegen Schmerzen in den Beinen immer wieder stehen zu bleiben. Unauffälliger und unterhaltsamer sind die Laufpausen vor Schaufenstern – daher der Name. Langfristig kann die Schaufensterkrankheit die Beinadern so beeinträchtigen, dass amputiert werden muss. In Deutschland werden jedes Jahr rund 30.000 Gliedmaßen wegen solcher arterieller Durchblutungsstörungen amputiert.

Ein neuartiges Medikament lässt die Betroffenen hoffen: Nach Angaben des Herstellers erhöht es die Laufleistung um sagenhafte 16 Prozent. Kein Wunder, dass er

dafür kämpft, dass das teure, innovative Mittel von den Krankenkassen übernommen wird. Zu Recht?

Als Journalist besorgen Sie sich die Unterlagen über die klinischen Tests, die Ihnen die Pharmafirma sehr zuvorkommend per Boten ins Haus schickt. Versiert, wie Sie sind, lesen Sie nach: Tatsächlich konnten die Patienten mit dem Medikament zwischen 6 und 26 Prozent längere Strecken zurücklegen als ohne. Der Fall scheint klar: Das Medikament wäre für die vielen Betroffenen ein Gewinn, die Krankenkassen verwehren aus purer Knauserigkeit ein innovatives Mittel.

Stimmt das? Bevor Sie das guten Gewissens schreiben dürfen, müssen Sie klären, um welche Größenordnungen es geht. Sie sollten deshalb einige Fragen klären, die Ihnen die Pharmafirma vermutlich nicht ganz so fix beantwortet. Die am nächsten liegende lautet: Auf welcher Grundlage sind die Prozentzahlen berechnet worden? Vergrößert sich der Aktionsradius der Betroffenen dank des neuen Mittels um 2.000 Meter oder um 10 Schritte?

Nächste Frage: Um wie viel ist das teurere Mittel besser als herkömmliche Mittel? Nehmen wir an, die etablierten Wirkstoffe gegen die Schaufensterkrankheit, deren Patentschutz lange abgelaufen ist und die es als günstige Nachahmerprodukte (Generika) zu kaufen gibt, verbessern die Laufleistung der Patienten um 10 Prozent. Damit schrumpft der Vorteil des neuen Mittels auf 6 Prozent zusammen.

Wenn ein Mensch ganz ohne Medikament nach 20 Schritten stehen bleiben muss, bis die Schmerzen in den Wadenmuskeln abklingen, heißt das: Er könnte es mit einem herkömmlichen Generikum 22 Schritte schaffen, mit dem teuren neuen Mittel 23,2 Schritte. Ein Schritt mehr bis zur Pause verändert die Lebensqualität des Patienten nicht und rechtfertigt damit sicher keine erhebliche Mehrbelastung des gebeutelten Gesundheitsbudgets.

Weil es keine bewiesene Wirkung ohne Nebenwirkung gibt, ist die Gabe eines Medikaments immer eine Abwägungsfrage. Weitere Fragen sind deshalb: Rechtfertigt die Wirkung eventuelle Risiken? Auch nach Wechselwirkungen und Langzeitfolgen sollte man fragen. Möglicherweise erhöht das Medikament das Risiko, dass später amputiert werden muss. Viele Patienten mit Durchblutungsproblemen leiden auch unter Bluthochdruck und nehmen deshalb Medikamente. Gibt es Wechselwirkungen mit diesen Mitteln?

Fehlt nur noch die Grundsatzfrage, ob Medikamente gegen Schaufensterkrankheit überhaupt das Mittel der Wahl sind. Die meisten Patienten sind Raucher. Vielleicht würde es mehr nützen, das Rauchen aufzugeben, und gewissenhaft ein Bewegungstraining zu absolvieren. Gibt es dazu Untersuchungen?

Am Ende der Recherche steht also nicht nur die Zahl, die ganz eindeutig für das Medikament spricht. Sie ist ein Mosaiksteinchen in der Beurteilung eines neuen Wirkstoffs, und wie sich herausstellt, nicht von der Farbe, die das Bild am Ende bestimmt. Wer das neue Medikament gegen seine Schaufensterkrankheit nimmt,

schaft es trotzdem nicht ohne Pause über eine vierspurige Straße. Der Nutzen im Alltag kann also vernachlässigt werden. Außerdem birgt der neue Wirkstoff möglicherweise Risiken – etwa für Komplikationen bis hin zur Amputation, zu Leberschäden oder verändert die Wirkung von Blutdrucksenkern. Zu guter Letzt ist das neue Produkt unvernünftig teuer. Die Krankenkassen haben also gute Gründe, die Kostenübernahme abzulehnen.

Risikoreduktion – absolut oder relativ?

Wer regelmäßig Olivenöl zu sich nimmt, senkt sein Krebsrisiko um 9 Prozent. Wer zusätzlich weniger rotes Fleisch und mehr Hülsenfrüchte isst, kann es sogar um 12 Prozent senken. Das berichtete die BBC, und alle druckten es nach. Weil solche Meldungen einen Nutzwert haben, werden sie gern in Magazinen und auf vermischten Seiten aufgegriffen. Frauenzeitschriften strotzen vor kleinen Texten, die vermelden, dass dies oder jenes Obst oder Gemüse ein Krankheitsrisiko um eine hübsche Zahl senke. Aber was heißt das? Wenn 100 Menschen neuerdings Olivenöl zum Salat essen, überleben dann neun mehr? Nein.

Dramatischer wird der laxer Umgang mit Prozentzahlen, wenn es um Nutzen von Arzneimitteln oder medizinischen Tests geht. Es ist essenziell, zwischen absolutem und relativem Risiko unterscheiden zu können. Meist wird das relative Risiko angegeben, weil das dramatischer klingt. Der Wert bezieht sich auf die Veränderung. Solange man nicht weiß, wie der Ausgangswert war, ist er aber nicht sehr aussagekräftig. Interessant ist das absolute Risiko, dessen Wert die tatsächliche Gefahr für jeden einzelnen beschreibt. Es wird selten angegeben, deshalb sollte man es berechnen können.

Beispiel: Krebs-Test

Urologen empfehlen ihren Patienten gern einen Bluttest, der auf Prostatakrebs hinweisen kann (PSA-Test). Der Test kostet rund 40 Euro und ist umstritten. Klarheit sollten zwei große Langzeitstudien bringen, die im *NEW ENGLAND JOURNAL OF MEDICINE* erschienen (Andriole 2009). Die erste Studie zeigt keinen Nutzen des Tests. Doch der Urologenverband jubelt: Durch den PSA-Test wür-

de die Zahl der Todesfälle um 20 Prozent gesenkt. Bei diesem recht überzeugenden Wert handelt es sich um die relative Risikoreduktion.

Für die zweite Studie wurden 162.000 Männer aus acht Ländern neun Jahre lang begleitet. Die Hälfte machte jährlich einen PSA-Test. Von 1.000 Männern in der Testgruppe starben in 9 Jahren 3 an einem Prostatakrebs, in der Vergleichsgruppe waren es 3,7.

Für die relative Risikoreduktion setzen wir die 3,7 als 100 Prozent. In der Testgruppe sind es 0,7 Prozent weniger. Das entspricht 18,9 Prozent.

Doch wie sieht das absolute Risiko für den Einzelnen aus? In der ungetesteten Gruppe stirbt in neun Jahren ein Mann (genauer 0,7) weniger als in der Testgruppe. Der Test rettet also einen von Tausend – das ist eine Risikoreduktion von einem Promille.

Wenn tausend Menschen neunmal einen Test machen, profitiert statistisch gesehen einer. 8.999 Tests werden also ohne jeden Nutzen, aber mit einigem Schaden gemacht: Unsichere Diagnosen machen Gewebeproben nötig. Wird ein Prostataumor tatsächlich erkannt, so ist auch die Therapie von zweifelhaftem Nutzen: 30 bis 70 Prozent der langsam wachsenden Tumoren würden unnötig therapiert, sagen Fachleute. Sie hätten nie Probleme verursacht, die Therapie sei allerdings nebenwirkungsreich und psychisch extrem belastend.

Beispiel: Cholesterinstudien

Die 4S-Studie gilt als Mutter aller Cholesterinstudien. Sie verspricht 30 % weniger Todesfälle durch Herz-Kreislauf-Erkrankungen dank Cholesterinsenkern. Die Zahlen:

- 4.444 Versuchspersonen (davon 82 % Männer und 18 % Frauen)
- 2.222 bekommen das Medikament (Verumgruppe), 2.222 ein Placebo
- 5,4 Jahre dauert die Untersuchung
- Todesfälle: 182 in der Verumgruppe, 256 in der Placebogruppe

Relative Risikoreduktion: $256 - 182 = 74$. 256 entspricht 100 %, 74 sind 29 %.

Absolute Risikoreduktion: 74 von 2.222 werden gerettet, das sind 3,3 % in 5,4 Jahren.

Das heißt: In jedem Jahr, in dem 1.000 Menschen den teuren und nebenwirkungsreichen Cholesterinsenker schlucken, profitiert knapp einer, genauer 0,6 Menschen.