



NZZ FOLIO

DIE ZEITSCHRIFT DER NEUEN ZÜRCHER ZEITUNG

Home	Aktuelles Heft	Nächstes Heft	Frühere Hefte	Newsletter
Feedback	Bestellung	Werbung	Impressum	Suchen

NZZ Folio 1/06

Geschönt, geschlampt, gelogen

Weil positive Studienergebnisse eher publiziert werden, nehmen es viele Forscher in der Medizin nicht so genau mit der Statistik.

Von Robert Matthews

Es gibt eine gute Nachricht für schwer gestresste Frauen: Laut einer Studie aus Dänemark liegt die Wahrscheinlichkeit, dass sie an Brustkrebs erkranken, um 40 Prozent unter dem Erkrankungsrisiko anderer Frauen. Und nun die schlechte Nachricht: Laut einer anderen Studie, die kürzlich von schwedischen Forschern durchgeführt wurde, ist ihr Brustkrebsrisiko doppelt so hoch.

Ja was nun? Zwei Studien, beide von angesehenen Forschern durchgeführt und in führenden medizinischen Zeitschriften veröffentlicht, und doch stehen die Ergebnisse in krassem Widerspruch? Das ist zwar verblüffend, aber keineswegs ungewöhnlich. Es vergeht kaum eine Woche, in der nicht eine Studie erscheint, deren Resultate früheren Forschungsergebnissen widersprechen: Starkstromleitungen und Leukämie, Salzkonsum und Bluthochdruck, Herzerkrankungen und Sport - die Resultate pendeln mal in diese, mal in jene Richtung, ohne je zu definitiven Erkenntnissen zu führen. Im Jahr 2002 machten zwei der angesehensten medizinischen Zeitschriften innert weniger Wochen mit zwei Studien über den Zusammenhang zwischen Rauchen und Brustkrebs Schlagzeilen. Die erste meinte, einen solchen Zusammenhang bewiesen zu haben, die zweite dementierte ihn rundweg.

Ganz ähnlich verhält es sich mit neuen Therapien: Kaum hat ein Forscherteam einen Durchbruch verkündet, kommt ein zweites Team des Wegs und wirft alles über den Haufen. In den frühen 1990er Jahren glaubte man, dass die Hormonersatztherapie das Risiko von Herzerkrankungen bei Frauen halbiere. 2002 bewies eine grossangelegte Studie, dass Hormonersatz keinerlei Nutzen bringt.

Was geht hier vor? Weshalb kommen so viele Studien zu

widersprüchlichen Ergebnissen? Das fragen nicht nur immer mehr Wissenschaftler, sondern auch eine verunsicherte Öffentlichkeit. Die Antwort lässt Zweifel an der Zuverlässigkeit medizinischer Forschungsergebnisse aufkommen, auch wenn sie in führenden Fachzeitschriften publiziert wurden.

Wissenschaftler stehen unter einem beträchtlichen Publikationsdruck, um ihre akademische Stellung und die Finanzierung ihrer Projekte zu sichern. Daher besteht die berechtigte Sorge, dass sie «ungünstige» Ergebnisse absichtlich unterschlagen und sich auf Resultate konzentrieren, die ihre Veröffentlichungschancen erhöhen. Ausserdem wird je länger, desto deutlicher, dass viele Forscher selbst die einfachsten Methoden, mit denen sich Fehler bei der Interpretation medizinischer Versuchsreihen eingrenzen liessen, nicht beherrschen.

Am beunruhigendsten dabei ist, dass die meisten Forscher bis heute auf statistische Methoden vertrauen, von denen man seit über 40 Jahren weiss, dass sie zu irreführenden Ergebnissen führen. Anstatt die Plausibilität eines Ergebnisses zu berücksichtigen, stellen sie stur auf die statistische Signifikanz ab, die schon durch ihre Definition in einem gewissen Anteil der Studien zu einem irreführenden Resultat führen muss (siehe Seite 53). Kein Wunder also, dass ein Grossteil der medizinischen Forschung, die heutzutage in Fachzeitschriften publiziert wird, im günstigsten Fall wenig stichhaltig, im schlimmsten Fall einfach falsch ist. Wie ihre Kollegen in anderen akademischen Disziplinen stehen auch die Mediziner seit einigen Jahren unter steigendem Publikationsdruck: Schreibe oder stirb, lautet das Motto. Aber erst in jüngster Zeit hat man damit begonnen, die Auswirkungen dieser Devise auf die Zuverlässigkeit der medizinischen Forschung zu untersuchen. Die Ergebnisse sind beunruhigend.

Ein Forscherteam unter Leitung von An-Wen Chan vom Centre for Statistics in Medicine, Oxford, hat die Originaldaten von über 100 veröffentlichten Berichten über klinische Versuchsreihen unter die Lupe genommen. Das Team suchte nach Hinweisen darauf, dass «ungünstige» negative Ergebnisse in den publizierten Artikeln weggelassen wurden, um die Chancen einer Veröffentlichung zu erhöhen. Bei über der Hälfte der überprüften Versuchsreihen stiessen die Statistiker auf erhebliche Diskrepanzen zwischen den ursprünglichen Zielen der Studie und den berichteten Resultaten, was die Vermutung erhärtet, dass die Forscher einfach ihre Daten nach publizierbarem Material durchkämmt haben. Daran scheint für Laien nichts Schlechtes zu sein, doch Statistiker wissen: Weil jedes signifikante Resultat mit geringer Wahrscheinlichkeit durch eine zufällige Verteilung der Messwerte zustande kommt, kann man letztlich in jedem Datenberg signifikante Resultate finden, wenn man nur lange genug danach sucht.

Das Team um Chan entdeckte ausserdem, dass schädliche Therapiewirkungen, die sich während der klinischen Testphase einstellten, oft nur unvollständig beschrieben wurden und dass zentrale Fragen wie etwa Schmerzintensität und Überlebensrate in den Berichten entweder vernachlässigt oder ganz weggelassen wurden.

In ihrem 2004 im «Journal of the American Medical Association» (JAMA) erschienenen Bericht weisen die Autoren darauf hin, dass man diese Auslassungen bei der Lektüre der veröffentlichten Aufsätze

schlechterdings nicht erkennen kann, und fordern deshalb die umfassende Offenlegung der Ziele und Ergebnisse jeder medizinischen Studie. Solche Forderungen scheinen besonders bei Forschungen angezeigt, die von der Industrie finanziert werden. 2003 wertete ein Team um Cary Gross von der Yale University School of Medicine über 1000 Studien daraufhin aus, ob sich ein Zusammenhang zwischen den Resultaten dieser Studien und ihrer Finanzierung zeige. Das Ergebnis: 80 Prozent der industriefinanzierten Forschungen kamen zu positiven Ergebnissen, während es bei unabhängigen Forschern nur knapp 50 Prozent waren. Man könnte versucht sein, diese Diskrepanz dadurch zu erklären, dass die Industrie einen besseren Riecher für wirksame Therapeutika hat. Aber das Team von Gross fand heraus, dass in industriefinanzierten Versuchsreihen die Arzneimittel oft mit Placebos oder mit schwachen Alternativpräparaten verglichen wurden, um möglichst eindruckliche Resultate zu erzielen. Der Effekt war gelegentlich dramatisch. So wiesen zum Beispiel über die Hälfte der industriefinanzierten Studien zu einem bestimmten Arzneimittel gegen Herzkrankheiten positive Ergebnisse aus, während exakt dieselbe Substanz in keiner einzigen der unabhängigen Studien für wirksam befunden wurde.

Die Erkenntnisse von Gross und seinen Kollegen wurden 2003 im «Journal of the American Medical Association» veröffentlicht. Die Autoren halten diese Diskrepanzen nicht zuletzt deshalb für besorgniserregend, weil der Anteil an industriefinanzierten Forschungen beständig steigt. Inzwischen werden rund zwei Drittel der klinischen biomedizinischen Forschung in den USA durch die Industrie unterstützt - das ist doppelt so viel wie 1980.

Um des Problems der selektiven Veröffentlichung positiver Resultate Herr zu werden, verlangen führende medizinische Fachzeitschriften von ihren Autoren, alle klinischen Versuche bereits in der Planungsphase zu registrieren. So soll verhindert werden, dass Versuche mit negativen Resultaten unter den Teppich gekehrt werden. Seit 2003 können solche Studien auch dem «Journal of Negative Results» zur Veröffentlichung unterbreitet werden.

Trotz dem Druck, der auf den Wissenschaftlern lastet, ist die selektive Berichterstattung von Forschungsergebnissen noch nicht zur Regel geworden. Die meisten Forscher fühlen sich nach wie vor verpflichtet, die Wahrheit über die Wirkung neuer Substanzen herauszufinden, wie auch immer sie aussieht. Aber sowenig man an ihren Motiven zweifeln darf, so berechtigt ist doch die Empörung über ihren Dilettantismus. Um es geradeheraus zu sagen: Viele Forscher begreifen die Methoden nicht, mit denen sie zu ihren Ergebnissen gelangen.

In meiner Analyse aller in einem Jahrgang der führenden Zeitschrift «Nature» erschienenen Aufsätze fand ich heraus, dass 20 Prozent der Autoren die statistischen Verfahren, die sie verwenden, nicht verstehen. Zu ganz ähnlichen Schlüssen kommt eine Studie der Universität Gerona in Spanien, die den beiden massgeblichen Forschungszeitschriften «Nature» und «British Medical Journal» eine Unzahl statistischer Fehler nachweisen konnte.

Obschon die meisten dieser Fehler trivial waren, wurden mehrere Prozent doch als so gravierend eingeschätzt, dass sie die Forschungsergebnisse verfälscht haben könnten. Diese Erkenntnisse sorgten zu Recht für

Empörung. Der «Economist» entrüstete sich über die «schlampige Statistik, die eine Schande für die Wissenschaft» sei.

Die Statistiker wiederum hat daran nur schockiert, dass andere Leute sich darüber aufregten. Hatten sie nicht schon seit Jahrzehnten die kläglichen statistischen Analysen moniert, mit denen Forscher selbst in den angesehensten Zeitschriften zu publizieren wagten?

In der medizinischen Forschung spielen statistische Methoden eine entscheidende Rolle. Man verwendet sie, um abzuschätzen, wie gross ein klinischer Versuch angelegt werden muss, um die Wirksamkeit einer neuen Substanz zu erweisen, aber auch, um die Überzeugungskraft der Ergebnisse einzuschätzen. Zumindest sollte es so sein. In Wirklichkeit versuchen die meisten Forscher einfach so viele Probanden zu bekommen, wie ihr Budget zulässt, und hoffen, die Versuchsgruppe sei hinreichend gross, um eine echte Wirkung zu erkennen. Sobald die Daten vorliegen, jagt man sie durch eine Statistiksoftware und hofft, dass dabei wenigstens ein «statistisch signifikantes» Ergebnis herausspringe, das zur Veröffentlichung in einer bedeutenden Fachzeitschrift taugt.

Das klingt alles ganz passabel, hat aber in der medizinischen Literatur zu einer Vielzahl von völlig irreführenden Ergebnissen geführt. Ein kurzer Blick auf die medizinischen Fachzeitschriften zeigt, dass an klinischen Forschungen selten mehr als ein paar hundert Patienten beteiligt sind. Das hört sich nach viel an, aber gemessen an den Anforderungen der theoretischen Statistik, sind diese Versuchsgruppen immer noch viel zu klein, um tatsächlich eine Wirkung nachzuweisen. Handkehrum neigen Forscher, die einfach nur darauf hoffen, dass sie genügend Patienten zusammenbekommen haben, beim Ausbleiben positiver Resultate dazu, die Therapie als wirkungslos zu verwerfen, während in Wahrheit nur die Versuchsgruppe zu klein war, um eine Wirkung nachzuweisen.

Die Wahrscheinlichkeit solcher «falsch negativen» Testergebnisse ist keineswegs zu vernachlässigen: Eine vergangenen August von John Ioannidis von der Universität Ioannina in Griechenland durchgeführte Untersuchung kam zu dem Schluss, dass drei Viertel aller kleinen Studien zu irreführenden Ergebnissen gelangen. Dieses Problem kommt besonders bei alternativen Therapien wie zum Beispiel der Akupunktur zum Tragen. Die Forscher, die sich mit solchen Therapieformen beschäftigen, verfügen meist nicht über dieselben Ressourcen wie die Schulmediziner, und häufig nehmen an ihren Studien weniger als 100 Patienten teil. Entsprechend gross ist das Risiko, dass dabei reale Wirkungen übersehen werden.

Die Erkenntnis, dass kleine Studien weniger zuverlässig sind als grosse, dürfte niemanden überraschen. Aber grosse Untersuchungen fallen leicht einem anderen statistischen Irrtum anheim, der für viele der widersprüchlichen Ergebnisse verantwortlich ist, wie etwa beim Zusammenhang zwischen Stress und Brustkrebs. Der Effekt ist seit Jahrzehnten bekannt, und führende Statistiker werden nicht müde, vor seinem Einfluss auf die Forschung zu warnen. Allein, ihr Rufen blieb weitgehend ungehört.

Um es einfach auszudrücken: Die statistischen Methoden, die routinemässig von den Forschern eingesetzt werden, lassen stets einen zentralen Faktor unberücksichtigt, der die Glaubwürdigkeit eines jeden Resultats beeinflusst: seine Plausibilität.

Wenn Wissenschaftler die Ergebnisse einer klinischen Versuchsreihe zur Wirksamkeit eines neuen Medikaments auswerten, benutzen sie Computerprogramme, die ihnen sagen, ob der Anteil von Patienten, deren Zustand sich nach Einnahme der Substanz verbesserte, wesentlich höher sei als bei alternativen Behandlungsmethoden. Sind die Unterschiede gering, kann es sich um einen Zufall handeln. Ist der Unterschied jedoch gross genug, sinkt die Wahrscheinlichkeit eines Zufallsergebnisses, und das Resultat gilt als «statistisch signifikant».

Dafür steigt die Wahrscheinlichkeit, dass solche Resultate in wichtigen Fachzeitschriften veröffentlicht werden. Aber die statistische Signifikanz sagt überhaupt nichts über die Plausibilität des behaupteten

Zusammenhangs aus. Sie verstösst vielmehr grundsätzlich gegen eine elementare Maxime jeder Wissenschaft: Je aussergewöhnlicher die Behauptung, desto aufwendiger muss die Beweisführung sein.

Es gibt verschiedene Methoden zur Berücksichtigung von Plausibilität, und wenn man diese Methoden auf die Ergebnisse klinischer Forschung anwendet, gelangt man zu erschütternden Erkenntnissen: Eine Vielzahl statistisch signifikanter Resultate erweisen sich als bedeutungslose Zufälle.

Dass es gefährlich ist, sich allein auf statistische Signifikanz zu verlassen, ist schon seit vielen Jahren bekannt. Dennoch entblöden sich selbst ernst zu nehmende medizinische Zeitschriften nicht, die abenteuerlichsten Behauptungen zu publizieren. Ein klassisches Beispiel erschien 2001 im «British Medical Journal». Es handelte sich um einen Aufsatz, der scheinbar zwingende Beweise für die medizinische Wirksamkeit von Gebeten erbrachte. Laut Studie stiegen die Heilungsaussichten in statistisch signifikantem Masse, wenn für die Patienten gebetet wurde - und sei es Jahre nachdem sie das Krankenhaus verlassen hatten! Das Ergebnis dieser Forschung legte nahe, dass Gebete auch rückwirkend helfen können, und prompt wurden Forderungen laut, unsere Vorstellungen über Zeit und Raum zu überdenken. In Wirklichkeit zeigt die Forschung nur, zu welch unsinnigen Schlüssen man kommt, wenn man bei der Auswertung neuer Daten ihre Plausibilität ausser Acht lässt. In den meisten Fällen sind die Gefahren jedoch längst nicht so offensichtlich. So werden seit einigen Jahren immer mehr Berichte über die Entdeckung von Genen veröffentlicht, die angeblich mit Krebs oder anderen Krankheiten im Zusammenhang stehen. All diese Behauptungen, die sofort Schlagzeilen machen, beruhen auf «statistisch signifikanten» Zusammenhängen zwischen der Anwesenheit dieser Gene und der Wahrscheinlichkeit, an einem bestimmten Leiden zu erkranken. Aber nur zu oft lösen sich diese Zusammenhänge in Luft auf, sobald andere Forscher sie zu bestätigen versuchen.

Das US National Cancer Research Institute in Bethesda führte im März 2004 die vorschnelle Proklamation solcher Zusammenhänge auf den übertriebenen Glauben an statistische Signifikanz zurück. Dasselbe Phänomen könnte für ein weiteres schockierendes Ergebnis verantwortlich sein, das im Juli 2005 publiziert wurde: Rund ein Drittel der häufig zitierten Resultate halten einer späteren Überprüfung nicht stand.

Sobald man jedoch die Plausibilität einbezieht, lassen sich viele der sonst verwirrenden Forschungsergebnisse erklären. Nehmen wir zum Beispiel

die beiden widersprüchlichen Studien über den Zusammenhang zwischen Stress und Brustkrebs. Laut der dänischen Studie haben Frauen, die regelmässig Stress ausgesetzt sind, ein statistisch signifikant geringeres Risiko, an Brustkrebs zu erkranken, wohingegen die schwedische Studie herausfand, dass das Krebsrisiko bei Stress statistisch signifikant erhöht ist.

Wenn man sich die Ergebnisse genauer ansieht, zeigt sich, dass die dänische Studie weit weniger aussagekräftig ist als die schwedische. Und was die Plausibilität angeht, muss erwähnt werden, dass ein Team der Harvard Medical School im Jahr 2000 die Ergebnisse einer riesigen Erhebung bei fast 27 000 Frauen veröffentlichte, die keinerlei Hinweis auf einen Zusammenhang zwischen Stress und Brustkrebsrisiko bei Frauen erbrachte. Eine so umfassende Studie hat ein entsprechend grosses Gewicht - im Vergleich dazu können beide skandinavischen Studien nicht überzeugen. Die Antwort auf die Frage, welcher von beiden Studien wir glauben sollen, lautet demnach: keiner! Ihre Ergebnisse mögen statistisch signifikant sein, glaubwürdig sind sie nicht.

Allmählich beginnen die massgeblichen medizinischen Zeitschriften die Gefahren zu erkennen, die sich durch die fehlerhafte Anwendung statistischer Methoden in die Forschung einschleichen. Dass sie bis heute nichts dagegen unternommen haben, ist ein Skandal. Und solange sich an ihrer Veröffentlichungspolitik nichts ändert, kann man nur jedem, der sich fragt, ob er brandneue, aber unwahrscheinliche Forschungsergebnisse ernst nehmen soll, nur raten: Vergiss es!

Robert Matthews ist Wissenschaftsjournalist und zurzeit Gastdozent an der naturwissenschaftlichen Fakultät der Aston University, Birmingham.

Übersetzung: Robin Cackett, Berlin.

Bücher zum Thema:



[Mit an Wahrscheinlich gr...](#)

Hans-Hermann Dubbe...

Bester Preis EUR 8,90
oder neu kaufen



[Information](#)



[Der Hund, der Eier legt](#)

Hans-Peter Beck-Bo...

Bester Preis EUR 6,90
oder neu kaufen EUR 9,90



[Information](#)



[Der Schein der Weisen](#)

Hans-Peter Beck-Bo...

Bester Preis EUR 2,85
oder neu kaufen EUR 8,90



[Information](#)

[[Aktuelles Heft](#) | [Nächstes Heft](#) | [Frühere Hefte](#) | [Feedback](#) | [Bestellung](#) | [Werbung](#) | [Impressum](#) | [Suchen](#)]



[[Neue Zürcher Zeitung](#) | [NZZ FOLIO](#) | [Format NZZ](#)]
