Schweiz Med Forum 2009;9(9):192–194

Statistik in der Onkologie

Chancen und Gefahren

Jan C. Schuller
Schweizerische Arbeitsgemeinschaft für klinische Krebsforschung SAKK, Bern

Summary

Statistics in oncology: opportunities and risks

● *The statistical quality of clinical reports is often questioned, the main problems being selective publication of favourable results, post-hoc and data-driven hypotheses, multiple testing and overinterpretation of p-values. One reason for this conflict lies in the different approaches taken by statistics and medicine; this includes education in the two fields, which is differently motivated. The attention of a medical doctor traditionally focuses on individual patients, whereas the statistician is interested in a hypothetical "average" patient.*

● *The present paper discusses possible causes for this conflict from the perspective of clinical oncology, with the aim of contributing to better interdisciplinary communication and thus improving the quality of clinical research.*

Einleitung

Oft wird die Qualität der statistischen Auswertung klinischer Forschungsdaten beklagt [1] und regelmässig über den kunstgerechten Gebrauch der Statistik bei der Planung, Auswertung und Publikation ebensolcher Studien diskutiert [2]. Die vorliegende Arbeit geht den möglichen Ursachen für diesen Konflikt aus Sicht der klinischen Onkologie nach. Die Kenntnis der Ursachen kann die interdisziplinäre Verständigung und damit die Qualität der klinischen Forschung verbessern.

Hauptkritik ist, die klinische Literatur sei eine Ansammlung positiver Ergebnisse [3]. Ein kürzlich im *New England Journal of Medicine* erschienener Artikel sah geradezu katastrophale Zustände bei der Subgruppenanalyse [4]. Es gibt eine ganze Reihe von Gegenmitteln, um den falschen Gebrauch der Statistik zu verhindern, angefangen bei der Korrektur von Signifikanzwerten bei multiplem Testen (z.B. Bonferroni-Korrektur) bis hin zu komplexen Massnahmen, wie der vorgängigen Registrierung von Studien als Vorraussetzung für deren Publikation, damit Transparenz darüber besteht, welches die Ausgangshypothesen waren.

Als Ursachen für die mitunter mangelhafte Statistik werden oft ökonomische Interessen in den Mittelpunkt gerückt, nach denen der klinische Versuch eher als Marketingmassnahme denn als wissenschaftliches Experiment verstanden wird. Darüber wird ein anderer Aspekt vernachlässigt, der für die Qualität einer Studie wenigstens ebenso bedeutsam ist. Denn bei der Durchführung von klinischen Studien trifft die Medizin auf die Statistik, und dies sind zwei Disziplinen, wie sie unterschiedlicher beinahe nicht sein könnten. Der Grundstein für eine korrekte statistische Auswertung einer klinischen Studie wird bereits in der Planungsphase gelegt. Hier kommt es zu einer intensiven Zusammenarbeit zwischen Statistikern und Klinikern. Diese hat zum Ziel, klinisch relevante Aspekte so mit einem statistischen Design zu verbinden, dass die Studie mit hoher Wahrscheinlichkeit ein eindeutiges Resultat liefert. Dieses Ziel wird umso schneller erreicht, je besser die Verständigung zwischen den beteiligten Fachpersonen funktioniert.

Auf diesen vernachlässigten Aspekt wird im Folgenden eingegangen und dabei ein besonderes Augenmerk auf die Ursachen der Kommunikationsprobleme zwischen Statistik und Klinik gelegt.

Ursachen

– Statistik und Medizin unterscheiden sich erheblich, was den wissenschaftlichen Ansatz und die Ausbildung der Vertreter beider Disziplinen angeht. Die Statistik ist ein Kind der Mathematik und agiert mit Axiomen und Beweisen. Die klinische Forschung hingegen braucht die Methoden der induktiven Naturwissenschaften. Die Ausgangsfrage des Arztes richtet sich gewöhnlich an das Individuum: «Wie heile ich den Patienten, der mir gegenüber sitzt?»
  Dem Statistiker gilt das Individuelle nur als Ausprägung einer verborgenen Wahrheit, die er zu ergründen hofft, indem er das gleiche Merkmal bei vielen Patienten misst. Der Statistiker benötigt für seine Arbeit mehr oder weniger homogene Patientenkollektive, auf die Standardmethoden angewendet werden können. Das individuelle Arzt-Patienten-Verhältnis befindet sich in gewissem Widerspruch dazu [5].
– Beide Disziplinen haben spezifische Fachsprachen, die den Dialog innerhalb des jeweiligen Fachgebiets erleichtern und das Zugehörigkeitsgefühl stärken. Gleichzeitig erhalten sie damit aber auch den Charakter von Geheimsprachen, deren Sinn die Abgrenzung zum anderen Fachgebiet ist und nicht die Verständigung.
– Statistik als Werkzeug: Eigentlich ist die Statistik ein Mittel zur Objektivierung von Sach-

verhalten. In der Praxis wird sie jedoch oft als Werkzeug gebraucht, um vorgefassten Meinungen einen mathematisch (d.h. exakt) anmutenden Anstrich zu geben. Die Methoden der Statistik mögen objektiv sein, ihre Anwendung ist es deshalb noch nicht. So gesehen besteht ein Widerspruch zwischen dem Konzept des *Key Opinion Leaders* und dem der *Evidence-Based Medicine*.

– «Für gute Ergebnisse braucht man keine Statistik» – diese Haltung demonstriert ein gewisses Misstrauen gegenüber der wissenschaftlichen Methode, so, wie Goethe dem Fernrohr misstraute, weil es die Wirklichkeit verzerrt. Die Statistik stellt gewissermassen ein Mikroskop dar, mit dessen Hilfe auch sehr kleine Effekte oder solche mit grosser Streuung sichtbar gemacht werden können. Am anderen Ende der Skala stehen die wissenschaftlichen Errungenschaften, deren Signifikanz einleuchtet, auch ohne Statistik (z.B. Penicillin oder Röntgenstrahlen).
  In der klinischen Onkologie hingegen können viele klinisch relevante Ergebnisse erst mit Hilfe der Statistik gewürdigt werden. Hier stellt sich die Frage, welche Strategie man bei klinischen Studien in der Onkologie verfolgt: Setzt man darauf, dass sich die vielen kleinen Resultate schliesslich zu einem grossen Effekt addieren, indem die Behandlungsmethoden aufgrund der Ergebnisse immer weiter verfeinert werden? Oder hofft man darauf, eines Tages einen «Durchbruch» zu erzielen? Falls Letzteres zutrifft, ist die Statistik tatsächlich überflüssig, da die gesuchten Effekte so gross sind, dass man sie auch mit blossem Auge, d.h. ohne Statistik, sehen kann.
– «Data Mining» bedeutet eine Art Schatzsuche nach verborgenen Perlen in der Datenflut. Wichtige Anwendungen finden sich in vielen Spezialgebieten, z.B. in der Bioinformatik. Die Hauptgefahr besteht darin, das Gefundene zu überschätzen. Denn in einer genügend grossen Datenmenge kann man Hinweise für praktisch jede beliebige Hypothese finden. Auch gibt es immer Daten, die interessant genug erscheinen, um nach einer «post hoc» geschneiderten Hypothese zu rufen. Das Problem ist die Multiplizität: Wenn man beim Poker vier Asse bekommt, beweist das zwar, dass sie im Kartenspiel vorhanden sind, aber nicht, dass man sie in der nächsten Runde wieder erhält. Manche Autoren hoffen vergeblich, das Problem der Multiplizität zu lösen, indem sie sehr viele Analysen vor dem Experiment spezifizieren. Die wichtigste Arbeit ist die Hypothesenbildung. Selbst die ausgefeilteste statistische Analytik kann diese nicht ersetzen.
– Die benötigte Patientenanzahl einer Studie ist der wichtigste Kostenfaktor und entscheidet über die Machbarkeit. Die Patientenanzahl (bzw. die Stichprobengrösse) ist auf der anderen Seite von der erhofften Grösse des nachzuweisenden Behandlungseffekts abhängig. Bei der klinischen Prüfung einer neuen Behandlung soll meist eine Verbesserung gegenüber der Standardmethode nachgewiesen werden – im Falle von Krebsbehandlungen ist dies in der Regel ein längeres Überleben der Patienten. Je kleiner der erwartete Überlebensvorteil durch die zu prüfende Behandlung ist, desto mehr Patienten werden benötigt, um den Effekt nachweisen zu können. So entsteht die paradoxe Situation, dass grosse Studien begonnen werden, um kleine, mitunter klinisch irrelevante Effekte nachzuweisen.
  Eine andere Gefahr droht, wenn der erwartete Behandlungseffekt unrealistisch gross eingeschätzt wird, um die Anzahl der benötigten Patienten zu verringern. Auch in diesem Fall werden Ressourcen für eine Studie verschwendet, die viel zu unempfindlich *(«underpowered»)* ist, um einen Behandlungseffekt nachzuweisen. Entsprechendes gilt für die Verwendung von sogenannten *Composite Endpoints*, die oft nichts anderes als ein künstliches Herunterrechnen der benötigten Patientenanzahl bedeutet [6].
  Bei der Beurteilung von Effektgrössen herrscht auch eine gewisse Verwirrung, weil diese mit Vorliebe als Anteile von Anteilen ausgedrückt werden. Eine Aussage wie «Das Zwei-Jahres-Überleben konnte um 20% verbessert werden» sagt jedoch wenig aus, wenn keine absoluten Zahlen für die Überlebenszeiten und -anteile geliefert werden.
– Randomisierte Phase-II-Studien: Mit Phase-II-Studien wird mit dem Einsatz einer relativ geringen Stichprobengrösse und nur einem Behandlungsarm die Vorabklärung unternommen, ob es sich lohnt, einen neuen Behandlungsansatz weiter zu erforschen. Falls eine neue Behandlung sich als vielversprechend erweist, kann sie in einer Phase-III-Studie mit einer Standardbehandlung verglichen werden. Die typische Stichprobengrösse in einer Phase-II-Studie beträgt in der Regel deutlich weniger als hundert, während eine Phase-III-Studie den Einschluss von mehreren hundert Patienten erfordern kann.
  Normalerweise ist eine Phase-II-Studie einarmig, aber viele glauben nun, mit einer zweiarmigen randomisierten Phase-II-Studie (ein Behandlungsarm, ein Kontrollarm) testen zu können, ob eine neue Behandlung besser wirkt als eine Kontrollbehandlung. Das ist ein Irrtum. Die relativ kleine Stichprobengrösse bei Phase II erkauft man sich mit der Annahme, das Ergebnis der Kontrollgruppe sei bereits bekannt. Selbst wenn man eine Phase-II-Studie mit einem Kontrollarm durchführt, spielt dessen Ergebnis bei der statistischen Beurteilung der neuen Behandlung keine Rolle,

weil für einen Vergleich die Stichprobengrösse zu klein ist. Daher ist der Einschluss eines randomisierten Kontrollarms bei einer Phase-II-Studie nur selten gerechtfertigt. Randomisierte Phase-II-Studien enthalten meist mehrere Behandlungsarme, von denen der wirksamste in einer nachfolgenden Phase-III-Studie geprüft wird. Erst hier wird die neue Behandlung gegen eine Kontrolle getestet. Das Missverständnis besteht hier in der Annahme, mit Hilfe der Statistik könne man die benötigten Stichprobengrössen kleinrechnen und trotzdem die gleiche Information gewinnen.

Fazit

Die klassische Auffassung besagt, dass medizinisches Wissen im Verlauf des «singulären Kolloquiums» zwischen Arzt und Patient entsteht. Klinische Studien, besonders multizentrische, geben dieser Singularität wenig Raum. Es kommt vielmehr darauf an, allgemeingültige Kategorien festzulegen und dann für jeden einzelnen Patienten eindeutig festzustellen, ob sie für ihn zutrifft oder nicht («A ist progredient, B ist es nicht»). Diese Vorgehensweise ist notwendig, um die gewonnene Information der statistischen Analyse zugänglich zu machen. Im günstigsten Fall kommt es zur Verarmung der Information – der Patient wird reduziert auf ein paar Zahlen, auf Grösse, Gewicht, Anzahl roter Blutkörperchen u.a.m.; im ungünstigen Fall droht die Verfälschung, dann nämlich, wenn es misslingt, einen Patienten aufgrund der gesammelten Daten zu «synthetisieren». Diese Gefahr besteht immer, wenn die Endpunkte einer Studie nicht korrekt operationalisiert sind oder Daten ad hoc interpretiert werden (also anders, als im Protokoll festgelegt).
Andererseits sind Intuition und Ad-hoc-Interpretation entscheidende Mittel der Erkenntnis. Allerdings ist manchmal schwer zu klären, wie man sie der statistischen Analyse zugänglich machen kann. Eine Abhilfe ist wiederum die genaue Festlegung der Endpunkte: Was will man messen, und welche Schlüsse zieht man daraus? Wenn dann aufgrund unerwarteter Ergebnisse unvorhergesehene Schlüsse gezogen werden, ist dies zumindest transparent.

Es ist überraschend, wie sehr forschende Mediziner (aber auch andere Naturwissenschaftler) bereit sind, statistischen Analysen, insbesondere signifikanten Tests, den Rang unbezweifelbarer Wahrheiten einzuräumen. Die Bedeutung von $p < 0{,}05$ kann man sich jedoch leicht am Beispiel von zwei Würfeln klarmachen: Zwei Sechser gibt es mit $p < 0{,}028$, d.h., man könnte bei zwei Sechsern mit $p < 0{,}05$ sagen, die Würfel sind gefälscht. Die individuelle Erfahrung sagt andererseits, dass zwei Sechser so selten nicht sind, auch mit fairen Würfeln. Daraus folgt, dass ein statistisch signifikantes Ergebnis einen durchaus erfahrbaren Grad von Unsicherheit aufweist. Zwar gibt es eine Unmenge an Lehrbüchern zu diesem Thema, in denen vom Wahrheitsgehalt statistischer Analysen die Rede ist, welcher ganz wesentlich vom Geist abhängt, in welchem sie durchgeführt werden, kritisch oder nur zur vermeintlichen Bestätigung eines Vorurteils. Die meisten Statistiklehrbücher für Nichtstatistiker, die oft Titel tragen wie «Statistik für Mediziner, Biologen oder Sozialwissenschaftler», sind allerdings hauptsächlich Rezeptsammlungen, bei denen die Grundlagen nur Beachtung finden, insofern sie für das Nachkochen unentbehrlich sind.
Andererseits sind diejenigen, die Statistik als eigenständige Wissenschaft betreiben, oft unbedarft, wenn es um Dinge der wirklichen Welt geht. Manchmal meint man, es fehle ihnen genau das, was einen guten Mediziner auszeichnet: die Intuition in der Analyse des Einzelfalls aufgrund langer praktischer Erfahrung.
So gesehen bilden Kliniker und Statistiker ein sich ideal ergänzendes Paar, wenn sie sich verständigen können und wollen. Unabdingbare Voraussetzung dafür ist grundlegendes Verständnis für das jeweils andere Fachgebiet. Die klinische Forschung kann davon nur profitieren.

Nachbemerkung

Die vorliegende Arbeit wurde anlässlich eines interdisziplinären Symposiums mit dem Titel *Statistics in Oncology* erstellt, welches die Schweizerische Arbeitsgemeinschaft für klinische Krebsforschung regelmässig durchführt.
Ich danke Chantal Britt, Shu-Fang Hsu Schmitz und Peter Brauchli für die wertvollen Diskussionen.

Korrespondenz:
Dr. sc. nat. Jan C. Schuller
SAKK
Schweizerische Arbeitsgemeinschaft für klinische Krebsforschung
Abteilung für Statistik
Effingerstrasse 40
CH-3008 Bern
jan.schuller@sakk.ch

Literatur

1 Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Medicine. 2005;2(8):e124.
2 Strasak AM ZQ, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research – a review of common pitfalls. Swiss Medical Weekly. 2007:44–9.
3 Dickerson K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990;263(10):1385–9.
4 Wang R LS, Ware JH, Hunter DJ, Drazen JM. Statistics in Medicine – Reporting of Subgroup Analyses in Clinical Trials. The New England Journal of Medicine. 2007;357(21):2189–94.
5 Helmchen H. Zwischen Individualisierung und Standardisierung. Deutsches Ärzteblatt online 2005:1–13.
6 Kleist P. Composite Endpoints: Proceed with Caution. Applied Clinical Trials online 2006(5).