AMB 1999, **33**, 25

Was lehren uns die großen Studien?

Zusammenfassung: Große Studien sind nicht nur für die Forschung, sondern auch für die therapeutische Praxis Erkenntnisquellen ersten Ranges. Sie beantworten jedoch immer nur eine oder wenige Fragen mit hoher Evidenz und decken keineswegs das gesamte Spektrum ärztlichen Handelns ab. Ihre Aussagekraft ist auch dadurch begrenzt, daß die weitreichenden Anforderungen an die Qualität klinischer Studien häufig nicht oder nur unzureichend erfüllt sind. Das Erkennen dieser Mängel ist oft nicht leicht und setzt das Vertrautsein mit spezifischen Begriffen zum Design von Studien und der Statistik voraus. Dennoch ist die Durchführung einer großen Studie in der Regel der einzige Weg, auf dem Klarheit in eine strittige Frage zum optimalen therapeutischen Vorgehen gebracht werden kann.

Traditionelle Erkenntnisquellen in der Medizin: Grundlage ärztlicher Kunst ist traditionell die auf der sorgfältigen Beobachtung des Einzelfalls fußende Erkenntnis über empfehlenswerte diagnostische Maßnahmen und therapeutische Interventionen; sie wird von Arzt zu Arzt und von Generation zu Generation persönlich kommuniziert und von Autoren von Lehrbüchern zu Deutungsmustern zusammengefaßt und kanonisiert.

Erst im Lauf der letzten hundert Jahre ist die Erkenntnis zum Allgemeingut geworden, daß die am Einzelfall orientierte Sichtweise zur Beantwortung einer Reihe wichtiger Fragen - insbesondere zur Bewertung der Zuverlässigkeit diagnostischer Verfahren und des Risiko-Nutzen-Verhältnisses therapeutischer Interventionen - in vielen Fällen nicht ausreicht. Während etwa die fiebersenkende Wirkung eines Medikamentes oder die antiinflammatorische Wirkung eines Antibiotikums bereits an wenigen Patienten evident werden kann, läßt sich z.B. die Reduktion kardialer Ereignisse durch eine cholesterinsenkende Diät bzw. eine medikamentöse Therapie oder die Frage der Lebensverlängerung bzw. -verkürzung durch Betreiben von Spitzensport nicht aus der persönlichen Erfahrung eines einzelnen Arztes oder durch den verbalen Austausch dieser Erfahrungen mit Fachkollegen zuverlässig beurteilen.

Hierzu sind wesentlich mehr Beobachtungen *(Fallzahlen)* erforderlich, die geordnet zusammengetragen und analysiert werden müssen.

Historisch schien diesem Bedürfnis nach größeren Fallzahlen zunächst die amtliche Medizinalstatistik Rechnung zu tragen. Es stellte sich jedoch heraus, daß derartige, hauptsächlich zu Verwaltungszwecken routinemäßig erhobenen Datensammlungen zum einen nicht den erforderlichen Detaillierungsgrad haben können, der zur Beantwortung spezifischer Fragen erforderlich ist, zum zweiten unterliegen sie charakteristischen Verfälschungen (z.B. *Vermengung von Effekten),* die Fehlinterpretationen wahrscheinlich machen. Beide Begrenzungen können nur durch geplante “große" Studien überwunden werden.

Große Studien: Das erste dieser beiden Probleme wird durch die *geplante epidemiologische Studie* aufgenommen. Da in diesen Studien zwar gezielt beobachtet, aber therapeutisch nicht interveniert wird, kann jedoch im Rahmen der Epidemiologie das Problem der vermengten Effekte grundsätzlich nicht gelöst werden. Auf die Frage, ob Raucher kürzer leben, weil sie rauchen oder weil sie auch sonst wenig gesundheitsbewußt sind, kann durch eine Beobachtungsstudie grundsätzlich nie endgültig geantwortet werden. Zwar erlauben statistische Kontrolltechniken in gewissem Umfang eine mathematische Korrektur bekannter und beobachteter Einflüsse; nicht zu beobachtende oder aus anderen Gründen nicht beobachtete Einflußfaktoren können aber jedes Ergebnis prinzipiell beliebig verfälschen, ja sogar ins Gegenteil verkehren. Das zweite Problem läßt sich daher grundsätzlich erst im Rahmen *randomisierter Studien* lösen; durch die *Zufallszuteilung* lassen sich zwar Irrtümer nicht grundsätzlich vermeiden, immerhin läßt sich aber die Wahrscheinlichkeit solcher Irrtümer auf ein kalkulierbares Maß begrenzen.

Die beschriebenen Erkenntnisquellen werden häufig zutreffend in Form einer *Evidenzkaskade* dargestellt: die höchste Stufe der Evidenz stellt demnach der sog. *“Goldstandard"* der randomisierten Studie dar, die niedrigste Form die klassische *Einzelfallzusammenstellung.* Geplante Studien stehen in ihrer Aussagekraft grundsätzlich über Sammlungen von Routinedaten. Zu beachten ist ferner, daß die Größe einer Studie nicht über ihren Platz in der Evidenzkaskade entscheidet. Eine kleine Studie ist in gewisser Weise sogar leichter “sauber" durchzuführen als eine

große Studie, insbesondere eine *Multicenter-Studie.* Hingegen sind große Studien dann unumgänglich, wenn die Studienergebnisse repräsentativ für unterschiedliche Randbedingungen sein sollen bzw. wenn feine Differenzierungen angestrebt werden.

Ärztliche Praxis: Nach dieser Beschreibung könnte man erwarten, daß geplante Studien als Erkenntnisquelle der modernen Medizin par excellence und letztlich als entscheidende Handlungsanleitung für die ärztliche Praxis anzusehen sind. Tatsächlich ist jedoch beim Praktiker häufig eine Reserve zu spüren, die Ergebnisse großer Studien schnell in die alltägliche Patientenbetreuung umzusetzen. Diese Skepsis ist zum Teil vernünftig aufgrund der im folgenden beschriebenen Begrenzungen großer Studien; zum Teil dürfte sie aber wohl nur auf mangelnden Kenntnissen über die Funktionsweise des Instrumentariums “Klinische Studie" beruhen.

Begrenzungen der Aussagekraft großer Studien: Die Aussagefähigkeit großer Studien wird eingeschränkt durch eine Reihe von Faktoren, die grob in zwei Klassen eingeteilt werden können: 1. Faktoren, welche die *interne Validität,* d.h. die Vergleichbarkeit der Teilpopulationen, gefährden, und 2. Faktoren, welche die externe *Validität,* d.h. die Übertragbarkeit der Ergebnisse auf andere Patientengruppen, gefährden.

Eine der wichtigsten Maßnahmen zur Sicherung der *internen Validität* einer Studie ist die maximal mögliche und vertretbare *“Verblindung"* der Studienteilnehmer. Am günstigsten ist es, wenn sowohl Patient wie Therapeut die Begleitumstände der Therapie beurteilen müssen, ohne die Identität der im Einzelfall angewandten Therapie zu kennen *(Doppeltblindstudie).* Damit werden unbewußte oder bewußte Präferenzen neutralisiert. In vielen Fällen scheitert allerdings die Verblindung an praktischen Hindernissen, etwa beim Vergleich einer medikamentösen mit einer interventionellen Therapie, selten aber auch an ethischen Hindernissen; selten deshalb, weil in der Regel der Schaden durch eine suboptimale Behandlung vieler zukünftiger Patienten höher zu bewerten ist als der Schaden durch eine möglicherweise suboptimale Behandlung der Studienpatienten. Für diesen Fall stehen jedoch Ersatztechniken, z.B. die Beurteilung des Therapieerfolges durch an der Therapie nicht beteiligte *“Blinded reader"* zur Verfügung. Wird dieser Punkt, die

Neutralisierung von Arzt- und Patientenerwartungen, in der Publikation einer Studie nicht speziell erörtert, so ist Skepsis bzgl. der Korrektheit der gezogenen Schlüsse angebracht.

Um zu prüfen, ob ein bei der Präsentation einer Studie vorgenommener Vergleich zweier Therapiegruppen tatsächlich fair war, können folgende Fragen hilfreich sein: Waren die Zugangswege zur Studie für beide Gruppen gleich? War die Begleitbehandlung gleich? Wurden beide Gruppen nach dem gleichen Schema auf Erfolg untersucht? Waren die Nachbeobachtungen vollständig? War die statistische Analyse fair (d.h. symmetrisch)?

Des weiteren ist Vorsicht bei der Interpretation eines Studienergebnisses angebracht, wenn es nicht gelungen ist, den Vorteil einer Therapie in der Gruppe aller für die Studientherapie vorgesehenen Patienten, der *Intention-to-treat-Population,* nachzuweisen, sondern statt dessen hierfür eine Auswahl aus dem Studienkollektiv erforderlich war. Die sogenannten *“Per-protocol"-Populationen* sind meist eine Auswahl besonders gut für die Therapie geeigneter Patienten und geben ein geschöntes Bild der klinischen Gesamtbilanz.

Eine weitere Verfälschungsquelle sind die in letzter Zeit immer häufiger gewordenen, ungeplanten Eingriffe des Studienmanagements in den Ablauf einer Studie während der Durchführungsphase. Der Wechsel von Endpunkten oder von Einschlußkriterien, der nicht statistisch abgesicherte Abbruch der Studie in einem für einen Wirksamkeitsnachweis günstigen Moment sowie die übereilte Vorlage “sensationeller" Studienbefunde geben Anlaß zur Frage, ob hier Ergebnisse geschönt und Kontrollgruppen benachteiligt worden sind. Ein hohes Gut ist in diesem Zusammenhang die *Prospektivität* einer Studie. Im Nachhinein ausgewählte und dann als Hauptergebnis ausgegebene statistische Analysen können fast beliebig verfälscht sein und sind in ihrer Validität kaum einzuschätzen. Mit etwas statistischem und klinischem Geschick lassen sich nämlich in jedem ausgedehnten Datenkörper interessante Effekte finden, die klinisch eindrucksvoll aufbereitet werden können. Solche *“Fishing expeditions"* bauschen häufig nur kleine oder sogar rein zufällige Effekte auf. Sie sind deshalb lediglich zum Erstellen von Hypothesen geeignet, die dann der unabhängigen Überprüfung bedürfen. Die Frage “Haben die Autoren die vorgelegten Analysen ex ante geplant?" ist daher eine der wichtigsten, um die Gültigkeit von Studienergebnissen zu beurteilen.

Immerhin kann man heute davon ausgehen, daß das Instrumentarium “Klinische Studie" einen Entwicklungsstand erreicht hat, der es wahrscheinlich macht, daß Mängel in der internen Validität einer Studie von fachkundigen Mitarbeitern der Zulassungsbehörden oder von einem *“Peer Reviewer"* einer besseren Zeitschrift gefunden werden. Ein Blick in die Bulletins und Editorials sowie das Verfolgen von Kongreßbeiträgen wird deshalb auch dem Praktiker ein gutes Bild davon vermitteln, wie es um die Verläßlichkeit des internen Vergleichs in einer konkreten Studie steht.

Demgegenüber ist die Beurteilung der *externen Validität* großer Studien außerordentlich schwierig. Die Schwierigkeiten beginnen damit, daß eine große Studie normalerweise Teil eines Entwicklungsprogramms ist und vor dem Hintergrund vieler vorausgegangener kleinerer Studien an speziellen Kollektiven unter restriktiven Bedingungen geplant worden ist. Auf der Basis dieser kleineren Studien - und damit auf vergleichsweise unsicherer Grundlage - sind aber bereits eine Reihe wichtiger Entscheidungen getroffen worden, die in der großen Studie gar nicht mehr untersucht werden, wie z.B. die genaue Indikation, die Applikationsform oder die Dosierung eines Medikaments. Das Studienergebnis der großen, entscheidenden Studie beinhaltet lediglich überzeugende Evidenz für die Hauptfragestellung, unter der die Studie geplant und statistisch optimiert wurde. Zu dieser Fragestellung gehören die gewählte Indikation und die Dosis als integraler Bestandteil. So läßt sich z.B. aus der WOS-Studie (s. AMB 1995, **29**, 92) lediglich lernen, daß die regelmäßige Einnahme von täglich 40 mg Pravastatin bei Patienten mittleren Alters mit mäßig erhöhten, diätresistenten Lipiden die Ereignisrate während 5 Jahren senkt. Damit wissen wir zwar mit einem hohen Evidenzgrad, daß diese vorgestellte Therapie tatsächlich “Ereignisse" verhindern kann, ohne gleichzeitig anderweitig wesentlich zu schaden. Ob das Ergebnis hingegen auf andere Patientengruppen und Dosierungen übertragen werden kann, ist in dieser Studie nicht untersucht worden; es ist letztlich bestenfalls Gegenstand von Plausibilitätsüberlegungen und schlimmstenfalls reine Spekulation.

Kritisch sind in diesem Zusammenhang auch alle Bemühungen der Studienverantwortlichen zu sehen, ein in ihrem Sinne ideales Studienkollektiv zu selektieren, etwa durch ausgedehnte Voruntersuchungen oder durch *“Run-in-*

Phasen". Der “Durchschnittspatient" oder die “Durchschnittstherapie" können sich im Einzelfall durchaus deutlich vom “typischen Studienpatienten" unterscheiden.

Schwer zu beurteilen ist im allgemeinen ferner die *innere Übertragbarkeit* einer Studie, d.h., ist der beobachtete Therapieeffekt wirklich allen Patienten in der Studie zugute gekommen? Gibt es Untergruppen, die nicht profitierten oder gar paradox reagierten? Einerseits erwartet man von einer großen Studie, daß sie die Reaktionen der Patienten in ihrer ganzen Bandbreite abbildet. Andererseits ist es selbst in großen Studien meistens praktisch unmöglich, Hinweise auf heterogene Reaktionen oder abweichende Untergruppen mit hinreichender statistischer Sicherheit zu beurteilen, um daraus therapeutische Konsequenzen zu ziehen. So ist die in der bereits zitierten WOS-Studie allgemein unterstellte Konstanz der Wirksamkeit in allen betrachteten Untergruppen lediglich eine nicht widerlegte Modellannahme und keineswegs eine statistisch gesicherte Hypothese. Bezüglich spezieller Fragestellungen in Subgruppen befindet sich auch eine große Studie ziemlich weit unten in der Evidenzkaskade. Gerade in diesem Punkt werden große Studien oft hemmungslos überinterpretiert. Für die sekundären Endpunkte von Studien gelten im übrigen ähnliche Überlegungen: für die einzelne Nebenzielgröße ist die Evidenz gering. Allerdings können sich in ihrer klinischen Bedeutung ähnlich zu interpretierende, aber unterschiedlich diagnostizierte Endpunkte gegenseitig stützen.

Bevor man ein Studienergebnis auf die eigene Praxis überträgt, ist ferner zu fragen, inwieweit die Studie von realistischen Therapiekonzepten ausgegangen ist, die unter den vor Ort gegebenen Verhältnissen überhaupt umgesetzt werden können. Insbesondere ist Zurückhaltung empfehlenswert, wenn eine therapeutische Intervention, wie meist in *Phase-II-Studien,* ausschließlich unter idealisierten Bedingungen (Spezialklinik) untersucht wurden.

Zur Beurteilung der Verläßlichkeit eines Ergebnisses ist es ferner empfehlenswert, sich den *Konfidenzbereich des Effektschätzers* anzuschauen. Er enthält die Bandbreite der Werte, die mit dem Studienergebnis bereits unter alleiniger Beachtung möglicher Zufallsschwankungen verträglich sind. Der schlechteste und der beste Fall liegen dabei häufig weit auseinander und demonstrieren die Unsicherheit, die auf der Basis gegenwärtigen Wissens noch besteht.

Zugang zu Ergebnissen klinischer Studien: Die Vielzahl möglicher kritischer Punkte an der Aussagekraft großer klinischer Studien könnte zu der Sichtweise verleiten, sie für wenig praxisrelevant zu halten. Das Gegenteil ist der Fall. Zurückhaltend interpretiert, sind die großen Studien in der Lage, Erkenntnisse über die Wirkungsweise von Therapien zu vermitteln, die in ähnlicher Klarheit auf keinem anderen Wege zu erhalten sind. Bereiche der Medizin, in denen große Studien schwer möglich oder wenig verbreitet sind (Hilfsmittel, Operationstechniken, Physiotherapie, Naturheilverfahren, weite Bereiche der Diagnostik), sind mit größeren therapeutischen Unsicherheiten und mit größerem Streit von “Schulen" behaftet als Bereiche, in denen große Studien fest etabliert sind (pharmakologische Forschung).

Leider ist der Zugang zu diesen Erkenntnissen nicht leicht. Um klinische Studien interpretieren und die Problematik der Interpretation verstehen zu können, muß man eine Reihe von epidemiologischen und statistischen Grundbegriffen in ihrer Essenz verstanden haben; sie werden leider in der traditionellen deutschen Medizinerausbildung nur unzureichend gelehrt. Zudem machen es die Autoren wissenschaftlicher Publikationen ihren Lesern nicht immer leicht, die Schwachpunkte und Begrenzungen ihrer Studien zu erkennen. Verständlicherweise identifizieren sie sich mit ihrer Studie und servieren dem Leser die kritischen Gegenargumente nicht gleich mit. Die Situation wird noch dadurch erschwert, daß häufig der erste Kontakt des Praktikers mit dem Ergebnis einer klinischen Studie nicht die Originalpublikation oder eine kritische Aufarbeitung durch unabhängige medizinische Zeitschriften ist, sondern der von Pharmareferenten mit überzeugendem Auftreten überreichte Flyer mit den schönsten Resultaten der Studie. Die von manchen unkritischen Apologeten der “Evidence based medicine" apostrophierte eigene Literaturrecherche im Internet ist als Gegenmittel nicht nur wegen des Zeitaufwands vor diesem Hintergrund unrealistisch. Die Ergebnisse großer Studien können nur im Diskurs mit von kritischem Geist beseelten Kollegen gewürdigt werden, wobei die Frage nach dem Wert und der Beweiskraft der Studie nicht nur zu ihrer Gesamt-, sondern jeder Teilaussage zu stellen ist. In diesem Sinne ist eine besondere Kultur des Umgangs mit klinischen Studien zu schaffen, die im Moment erst in Ansätzen zu erahnen ist, aber alltäglich werden sollte.

Dennoch sollte man von den großen Studien nicht die Beantwortung aller oder auch nur der meisten Fragen erwarten, die den Arzt in seiner täglichen Praxis bedrängen. Letztlich können Studien immer nur die grobe Richtung therapeutischen Handelns vorgeben. Die Anpassung an den einzelnen Patienten erfolgt in einem filigranen Prozeß gegenseitiger Steuerung, in dem Arzt und Patient stetig neue Impulse setzen und aus der Antwort des anderen ihre Schlüsse ziehen. Diesen komplizierten Wechselwirkungsprozeß im Zusammenleben zwischen Arzt und Patient können klinische Studien ohnehin nur grob vereinfacht darstellen. In der Statistik sind diese Anpassungsprozesse und die Entwicklung von Krankheitsvorgängen, die ja zeitlich ablaufen, nur gemittelt dargestellt, es fehlt die Dynamik. Die Betreuung von Patienten erstreckt sich überdies häufig über einen Zeitraum, der den Horizont der meisten großen Studien sprengt. Für die Arbeit im Feld haben daher die eingangs erwähnten traditionellen Vorgehensweisen aus der Prästudienära weiterhin Bedeutung - modifiziert im Schlaglicht der punktuellen Erkenntnisse, die uns die großen Studien vermitteln.

Empfehlenswerte Literatur

1. User's guide to the medical literature
I. How to get started.
Oxman, A.D., et al.: JAMA 1993, **270**, 2093.

II A. How to use an article about therapy or prevention.
Are the results of the study valid?
Guyatt, G.H., et al.: JAMA 1993, **270**, 2598.

II B. How to use an article about therapy or prevention.
What were the results and will they help me in caring for my patients?
Guyatt, G.H., et al.: JAMA 1994, **271**, 59.

III A. How to use an article about a diagnostic test.
Are the results of the study valid?
Jaeschke, R., et al.: JAMA 1994, **271**, 389.

III B. How to use an article about a diagnostic test.
What are the results and will they help me in caring for my patients?
Jaeschke, R., et al.: JAMA 1994, 271, 703.

IV. How to use an article about harm.
Levine, M., et al.: JAMA 1994, **271**, 1615.

V. How to use an article about prognosis.
Laupacis, A., et al.: JAMA 1994, **272**, 234.

VI. How to use an overview.
Oxman, A.D., et al.: JAMA 1994, **272**, 1367.

2. A consumers guide to subgroup analyses.
Oxman, A.D., et al.: Ann. Intern. Med. 1992, **116**, 78.

3. Surrogate end points in clinicals trials: are we being misled?
Fleming, T.R., und DeMets, D.L.: Ann. Intern. Med. 1996, **125**, 605.