Bundesgesundheitsbl -
Gesundheitsforsch - Gesundheitsschutz
2001 · 44:876–882 © Springer-Verlag 2001

B. Gibis · C. Gawlik · Kassenärztliche Bundesvereinigung, Köln

Hierarchie der Evidenz

Die unterschiedliche Aussagekraft wissenschaftlicher Untersuchungen

Zusammenfassung

In der medizinischen Erkenntnistherorie hat mit der Einführung der randomisierten, kontrollierten Studie in den vierziger Jahren ein Paradigmenwechsel stattgefunden. Nach der Abkehr von der retrospektiven, individuellen Erfahrung Einzelner als Maßstab für die Wirksamkeit einer Methode gilt für die Etablierung der Wirksamkeit die randomisierte kontrollierte Studie (RCT) derzeit als Goldstandard. Entsprechend der Fragestellung (Therapie, Diagnostik, Prävention, Klärung von Nebenwirkungen), des Problemhintergrundes (z. B. Erkrankung, gegenwärtige therapeutische Alternativen) und den Charakteristika der zu untersuchenden Intervention (z. B. Medikament, Operation, bildgebende Diagnostik) kann aus einem reichhaltigen Instrumentarium der klinischen Forschung das entsprechend angemessene Studienkonzept ausgewählt werden. Dabei lassen sich nicht immer alle Möglichkeiten des Erkenntnisgewinns ausschöpfen: Sichergestellt werden sollte hingegen immer, dass den Patienten die bestmögliche Information und auch Therapie zur Verfügung gestellt wird. Ausgehend von dem Goldstandard RCT hat sich deshalb eine Hierarchie der Evidenz etabliert, die verschiedenen klinischen Erkenntnissen (von der Expertenaussage bis zur klinischen Studie) unterschiedliche Aussagekraft beimisst. In der täglichen Praxis findet diese Evidenzhierarchie Anwendung sowohl in der evidenzbasierten Medizin, bei der Erstellung von HTA-Berichten als auch in der Versorgungsplanung wie beispielsweise im Bundesausschuss der Ärzte und Krankenkassen und hat sich dort bewährt.

Schlüsselwörter

Evidenzhierarchie · Randomisierte, kontrollierte Studie · Erkenntnistheorie · Bundesausschuss der Ärzte und Krankenkassen

Im folgenden Artikel sollen die gegenwärtigen Grundsätze des Erkenntnisgewinns in der Medizin beleuchtet werden. Zunächst werden die historischen Entwicklungen des Erkenntnisgewinns in der Medizin kursorisch dargestellt. Anschließend werden am Beispiel des Wirksamkeitsnachweises in Therapiestudien die aktuellen Prinzipien des Erkenntnisgewinns erläutert, darauf aufbauend wird auf die so genannte „Hierarchie der Evidenz" eingegangen und schließlich die Probleme der klinischen Forschung angesprochen. Als Sonderfall wird zudem die Diagnostik kurz gestreift und die festgestellten Ergebnisse in einem Fazit für die Praxis zusammengefasst.

Erkenntnisgewinn in der Medizin

Die Frage, ob eine medizinische Maßnahme, sei sie eine therapeutische, eine vorbeugende oder diagnostische, wirkt oder nicht, begleitet seit jeher die Menschen in ihren unterschiedlichen Rollen als Patient, als Heilkundler oder auch als Gesunde, die einen Krankheitsfall beobachten. Beeinflusst durch Spontanverlauf, konkurrierende Risiken oder auch unbeabsichtigte Verhaltensänderungen des Patienten ist jeder Krankheitsverlauf – auch ohne Einflussnahme eines Therapeuten – durch eine Veränderung zum positiven wie zum negativen geprägt. Während Verbesserungen gerne auf die vorher ergriffenen Maßnahmen zurückgeführt werden, sind Verschlechterungen aus Sicht der Therapeuten durch die fehlende Compliance des Patienten, das nicht Vorhandensein von therapeutischen Optionen oder „höhere Kräfte" bedingt. Dieser Sichtweise widersprechend reduzierte Voltaire etwas despektierlich die Rolle des Arztes mitsamt der angewendeten Interventionen auf den früher noch weniger als heute beeinflussbaren Krankheitsverlauf auf die Aussage: „Der Arzt hat den Patienten so lange zu unterhalten, bis die Natur ihre Wirkung zeitigt."

Nachdem jahrtausendelang unter einer Vielzahl von Maßnahmen nur wenig wirksame Interventionen enthalten waren (und diese nicht als solche erkannt wurden), war das Kriterium des (selten über anekdotische Berichte hinausgehenden) therapeutischen Erfolges nur von mittelbarer Bedeutung für die Anwendung einer Methode. Wichtiger noch war die zeitgeschichtliche Plausibilität der ergriffenen Maßnahme, mit anderen Worten: Die Möglichkeit der Wirksamkeit musste aus der Sicht des Arztes und zu einem gewissen Grad auch aus Sicht des Patienten plausibel sein. Unschuld postuliert hier [1]: „Die

Dr. Bernhard Gibis
Kassenärztliche Bundesvereinigung,
Herbert Lewinstraße 3, 50935 Köln;
E-mail: evaluation@kbv.de

Bundesgesundheitsbl -
Gesundheitsforsch - Gesundheitsschutz
2001 · 44:876–882 © Springer-Verlag 2001

B. Gibis · C. Gawlik

Hierarchy of evidence. About the validity of study designs in medicine

Abstract

Randomized controlled trials (RCTs) were introduced into clinical research in the fourties, in the United Kingdom. The RCT established as a gold standard that is still the best option for proving efficacy of medical interventions. However, as well as testing efficacy, clinical trials must analyze the safety of therapeutic methods. Retrospective, individual experiences or case series remain important in medical science, but, due to methodological flaws, both are regarded as less reliable when compared to evidences obtained from prospective, controlled trials. Depending on the method used for analyzing efficacy the information obtained differ in quality. Therefore a hierarchy of evidences has been established. Initially it has been used to assess the quality of preventive interventions, but it may also be applied to other fields, for example, diagnostics. This hierarchy of evidence is one of the cornerstones of evidence-based medicine. It is also of value in areas such as accreditation and approval of medical interventions. For instance, the statutory German Standing Committee of Physicians and Sickness Funds has adopted a hierarchy of evidence for the accreditation of medical interventions in public, ambulatory health care.

Keywords

Medicine, evidence-based · Randomized controlled trial · Hierarchy of evidence · Health care system

Akzeptanz eines Systems heilkundlicher Ideen und Praktiken in einer Bevölkerung ist in erster Linie der Überzeugungskraft der Ideen zuzuschreiben: therapeutische Erfolge in der klinischen Praxis sind nur von zweitrangiger Bedeutung". Und weitergehend: Außermedizinische Faktoren waren ebenso wichtig, wenn nicht gar wichtiger als der therapeutische Erfolg. Hierunter fallen beispielsweise das sozioökonomische Umfeld, gesellschaftspolitische Theorien, Grundwerte oder Ängste, die einen Zeitgeist widerspiegeln, der Einfluss nimmt auf die Art und Weise, wie „die Bedrohung durch Kranksein und frühen Tod empfunden, konzeptualisiert, in Theorien eingefügt und schließlich als heilkundliches Ideensystem kulturell strukturiert wird".

„Als Beleg für die Wirksamkeit einer medizinischen Maßnahme gilt die individuelle Erfahrung des Arztes."

Verstärkt und in Einklang gebracht mit diesem Plausibilitätsansatz wird die schon immer gemachte individuelle und in der Regel retrospektive Erfahrung des Arztes. Diese war von Hippokrates über Galen, Avicenna und van Swieten oder Sydenham wesentlicher Bestandteil der wissenschaftlichen Arbeit, von Martini „Deskription" genannt, die oft einer „genialen Beschreibung" [2] des beobachteten Sachverhaltes gleichkam und gleichsam als Beleg für die Wirksamkeit einer Intervention angesehen wurde. Dass diese Erfahrung jedoch fehleranfällig sein kann, stellt schon Wunderlich bei seiner Antrittsvorlesung über den „Plan zur festeren Begründung der therapeutischen Erfahrungen" bei der Übernahme des Leipziger Lehrstuhls für Innere Medizin am 12. März 1851 fest: „Die gewöhnlich einzige Gewähr für den Erfolg einer Behandlungsmethode sind die Versicherungen aus den Reminiszenzen der Praxis. Es ist schon schlimm, wenn die therapeutische Erfahrung des Einzelnen auf nichts als auf Reminiszenzen des Selbsterlebten aufgebaut ist; denn man weiß, wie trügerisch diese Erinnerungen sind, wie gerade die auffallenden, exzeptionellen Fälle am meisten sich einprägen, wie gern die Fälle im Gehirn sich mit der Zeit verdoppeln und verdreifachen, und wie es auf die subjektive Stimmung ankommt, ob man die Erfahrung häufig oder selten gemacht zu haben glaubt. Was für den Vorsichtigen manchmal heißt, das ist für den Sanguiniker oft oder immer, für den Zweifler selten oder niemals. Was soll aber daraus werden, wenn widerstreitende Behauptungen auf individuelle Reminiszenzen gestützt einander gegenüberstehen: Wie soll da jemals eine Entscheidung möglich werden?"

Im 20. Jahrhundert kam es nicht zuletzt basierend auf den Entwicklungen der angewandten Statistik (siehe z. B. [3]) zu einer rasanten Fortentwicklung der Methodik des medizinischen Erkenntnisgewinns in klinischen Studien. Grundlage war ebenso eine umfassende Erkenntnistheorie, so wie sie vor allem von Popper in den Jahren nach dem 2. Weltkrieg entscheidend geprägt wurden [4].

Therapiestudien

Um sich über die Effekte einer medizinischen Intervention Gewissheit zu verschaffen, drängen sich drei Fragestellungen auf [5]:

1. Wie verlässlich wurde der Effekt beobachtet?
2. Wie groß ist der beobachtete Effekt? Und schließlich
3. welche klinische Wichtigkeit hat er?

Unter medizinischer Intervention werden in diesem Zusammenhang sämtliche Maßnahmen und Initiativen verstanden. So fällt die Wirkung eines Medikamentes oder eines Medizingerätes genauso darunter wie der chirurgische Eingriff zur Operation eines Leistenbruches sowie ein verändertes Versorgungskonzept (z. B.: Fußambulanzen zur Versorgung diabetischer Wunden). Neben der Wirksamkeit spiegeln Begriffe wie Wirkung, Nutzen, Verträglichkeit oder Effektivität weitere Aspekte wider, ohne dass es zu diesen Begriffen allgemeinverbindliche Definitionen gäbe. Für den Nachweis der Wirksamkeit hat sich im Unterschied zu den anderen erwähnten Kriterien in den letzten Jahren ein Standardinstrumentarium zur Planung, Durchführung und Auswertung von klinischen Studien herausgebildet, das zunächst besonders in der Arzneimittelforschung Anwendung fand und zunehmend auf andere Bereiche der Medizin übertragen wurde (z. B. [6]). Zwei Zielen

Übersicht 1
Häufige Fehlertypen in klinischen Studien

Zufallsfehler

Bias (z. B. Selection-Bias, Recall-Bias)

Confounding (z. B. Alter, Geschlecht, sozioökonomischer Status)

Effekt-Modifizierung

dient dieses Instrumentarium: dem Schutze der Patienten (und/oder Probanden) in der klinischen Forschung (explizit definiert nach den Nürnberger Prozessen in der Deklaration von Helsinki und weiterentwickelt in den GCP[1]- und neuerdings ICH[2]-Richtlinien zur klinischen Forschung) [7, 8] und dem möglichst irrtumfreien Erkennen eines Therapieeffektes [9].

Um diese Ziele zu erreichen, wurden verschiedene Studienkonzepte entwickelt, die mehr oder weniger zuverlässig Therapieeffekte abbilden können. Zwar ist kein Studienkonzept völlig frei von möglichen Irrtümern, verzerrenden Einflüssen oder zufälligen Ergebnissen, doch sind bestimmte Studienkonzepte besonders anfällig, falsche oder fehlerhafte Resultate zu liefern. Im Weiteren werden Details – wenn auch nicht erschöpfend – dieser Konzepte erläutert, dabei wird insbesondere auf den Wirksamkeitsnachweis abgehoben.

Studienergebnisse können mannigfaltig beeinflusst werden: Werden zu wenige Patienten beobachtet, kann das festgestellte Ergebnis zufällig zustande gekommen sein (Zufallsfehler, s. Übersicht 1). Einflussgrößen, die unerkannt den Krankheitsverlauf mitbestimmen, können die Wirksamkeit einer Intervention vortäuschen (sog. Confounder wie z. B. das Lebensalter). Der Untersucher kann das Ergebnis beeinflussen, indem er Patienten auswählt, bei denen ein günstiger Spontanverlauf der Erkrankung zu erwarten ist (sog. Auswahl-Bias).

Bei der Befunderhebung nach Behandlung kann durch gezielte Fragestellung ein entsprechend positiver Befund induziert werden (sog. Auswerter-Bias). Oft bemühen sich Patienten, die sich von ihrem Arzt für die Teilnahme an einer Studie haben überzeugen lassen, dem gewünschten Ergebnis gerecht zu werden (sog. Hawthorne-Effekt). Ob bewusst oder unbewusst, es besteht prinzipiell die Gefahr, dass ein Untersucher, der von der Sinnhaftigkeit der untersuchten Maßnahme überzeugt ist, die Studienplanung und Studienauswertung in Richtung eines günstigen Ergebnisses beeinflusst.

Viele dieser Fehlerquellen betreffen alle Studientypen, manche sind jedoch symptomatisch für so genannte unkontrollierte Studien (Fallserien, Kohortenstudien ohne Kontrollgruppe, Kasuistiken, Katamnesen), bei denen nur eine Gruppe von Patienten beobachtet und ausgewertet wird (Tabelle 1). Hier kann nur bedingt der Spontanverlauf als Besserungsursache ausgeschlossen werden. Mit anderen Worten: ob der Behandlungseffekt trotz oder wegen der Behandlung eintrat, ist nicht sicher zu klären. Auch kann nicht geklärt werden, ob der erzielte Behandlungseffekt auf unspezifische Faktoren wie Zuwendung, Eindruck auf den Patienten u.ä. zurückzuführen ist (s. auch in einer sehr interessanten Zusammenstellung [10]).

Kontrollieren lassen sich diese Fehler derzeit nur in vergleichenden Studien. Zwei oder mehrere Gruppen von Probanden/Patienten sollen die gleichen Risiken aufweisen, die gleiche Standardbehandlung erfahren und auch sonst in allen Charakteristika möglichst identisch sein und im Idealfall nur eine Unterscheidung aufweisen: die Intervention (neues Medikament, neues Operationsverfahren etc.). Zu einem oder mehreren vorab festgelegten Zeitpunkten werden beide Gruppen untersucht

Tabelle 1
Therapiestudientypen

Studiendesign	Protokoll	Wirksamkeitsnachweis möglich?
Vergleichende Studien		
Randomisierte, kontrollierte Studie	Teilnehmer werden nach dem Zufallsprinzip entweder der Interventions- oder der Kontrollgruppe (Placebo, aktive Kontrolle etc.) zugeteilt, die Ergebnisse beider Gruppen werden miteinander verglichen.	ja
Vergleichende Studie, jedoch nicht randomisiert	Gruppenbildung wie oben, jedoch nicht nach dem Zufallsprinzip (z. B. Zuteilung nach Geburtsdatum, Reihenfolge der stationären Aufnahme etc.).	nein
Kohortenstudien	In der Regel prospektive Beobachtung einer definierten Gruppe von Patienten über einen definierten Zeitraum, mit gleichzeitiger (kontemporärer) oder historischer Kontrollgruppe oder ohne Kontrollgruppe.	nein
Fallkontrollstudien	Studienteilnehmer mit der zu untersuchenden Erkrankung oder Zielkondition, Kontrollpatienten werden „gematcht". In der klinischen Forschung seltener angewendeter Studientyp.	nein
Studien ohne Vergleich		
Nicht vergleichende Studien, Fallserien, Anwendungsbeobachtungen	Eine Gruppe von Patienten wird über einen Zeitraum beobachtet und Vorher-Nachher-Vergleiche innerhalb dieser Gruppe durchgeführt.	nein
Kasuistiken, Katamnesen	Einzelfallanalysen einzelner Patienten oder Krankengeschichten	nein

[1] Good clinical practice guidelines, inzwischen Teil der ICH-Guidelines.

[2] International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use (s. auch http://www.ifpma.org/ich1.html)

Tabelle 2
Levels of Evidence für Therapiestudien des Oxford Center for Evidence based Medicine (deutsche Übersetzung nach [26])

Empfehlungsgrad	Evidenzstufe	Evidenzgewinn bei Therapie oder Intervention
A	Ia	Systematische Übersicht von RCTs mit Homogenität der Ergebnisse
	Ib	Einzelner RCT mit engem Konfidenzintervall
B	IIa	Systematische Übersicht von Kohortenstudien mit Homogenität der Ergebnisse
	IIb	Einzelne Kohortenstudie oder RCT mit deutlichen Qualitätsmängeln (z. B. <80% Nachbeobachtung)
	IIIa	Ergebnisse aus Outcome-Forschung
	IIIb	Einzelne Fallkontroll-Studie
C	IV	Fallserien oder Fallkontroll- und Kohortenstudien mit deutlichen Qualitätsmängeln
D	V	Expertenmeinung ohne Beleg durch kritisch bewertete Literatur, auf der Grundlage von physiologischen Beobachtungen oder Laborforschung

und die Behandlungsergebnisse miteinander verglichen. Die Vergleichsgruppen werden je nach Fragestellung gewählt: Geht es um die grundsätzliche spezifische Wirksamkeit oder um Risiken der Intervention, ist, falls möglich, eine Placebokontrolle vorzuziehen. Ist eine placebokontrollierte Studie aus ethischen Gesichtspunkten (z. B. eine andere nachgewiesen wirksame Behandlung bei lebensbedrohlichem oder die Lebensumstände elementar einschränkendem Krankheitsverlauf steht zur Verfügung) nicht möglich oder steht der Vergleich mit einem Standardverfahren im Vordergrund, ist beispielsweise der Vergleich mit dem Verlauf unter Standardtherapie möglich. Grundlegende Bedeutung hat die Gleichverteilung aller bekannten und auch unbekannten Risiken und Faktoren in den Vergleichsgruppen zu Beginn der Studie. Dem Ziel einer möglichst hohen Annäherung an diese Gleichverteilung kommt derzeit die Zufallszuordnung zu den Therapiegruppen, die sog. Randomisation, am nächsten. Andere Verfahren (historische Vergleichsgruppe, konsekutive Zuteilung in Studiengruppen etc.) bergen prinzipiell die Gefahr, dass z. B. der Schweregrad der Erkrankung ungleich verteilt ist und damit der Heilungsverlauf schon per se in beiden Gruppen unterschiedlich ist.

Bewussten und unbewussten Beeinflussungsmöglichkeiten durch den Auswerter lässt sich durch geeignete Instrumentarien begegnen. Beispielsweise sollte die Randomisation so gewählt werden, dass eine Beeinflussung durch den Studienarzt nicht möglich ist (sog. concealment). Um dies sicherzustellen, wird die Telefonrandomisation empfohlen, bei der der Studienarzt in einem Zentrum anruft, das ihm nach dem Zufallsverfahren mitteilt, in welche Gruppe der Patient aufgenommen werden soll [11]. Wenn es um die Klärung spezifischer Wirkanteile einer Intervention geht, steht die Möglichkeit der Verblindung zur Verfügung. Im Idealfall wissen weder der behandelnde Arzt noch der Patient, welche Behandlungsmethode angewendet wird (Placebokontrolle). Diese Form der Verblindung ist bei Medikamenten häufig, bei anderen ärztlichen Methoden, wie z. B. der Operation, nur eingeschränkt oder gar nicht möglich (gleichwohl wurden verblindete randomisierte Studien auch in der Chirurgie durchgeführt, siehe als aktuelles Beispiel [6]). Ziel ist deshalb nicht die prinzipielle Durchführung der doppelblinden Studie, sondern entsprechend der Fragestellung und der zu untersuchenden Intervention die mit entsprechendem Instrumentarium ausgestattete kontrollierte Untersuchung.

Die vorangehend aufgeführten Beispiele lassen erkennen, dass in der klinischen Forschung derzeit die randomisierte, kontrollierte Studie (RCT) als das Optimum hinsichtlich des Wirksamkeitsnachweises angesehen wird. Andere, wenn auch häufig verwendete Studienformen, sind schon aufgrund ihrer Anlage so mit einem Fehlerpotenzial behaftet, dass die gewonnenen Ergebnisse sehr zurückhaltend interpretiert werden müssen und häufig allenfalls zur Generierung von Hypothesen geeignet sind, also allenfalls Annahmen bezüglich des wahren Sachverhalts ermöglichen [5]. Außerhalb dieser erkenntnistheoretisch begründeten Forderung nach der kontrollierten, vergleichenden Erfahrung existieren derzeit keine kohärenten Erkenntnismodelle (beispielsweise in der Komplementärmedizin), die einen validen Wirksamkeitsnachweis ermöglichen würden.

Evidenzhierarchie

Die im vorstehenden Abschnitt erläuterten Unterschiede in der Aussagekraft unterschiedlicher Studienkonzepte bilden auch die Basis der kritischen Beurteilung klinischer Studien (critical appraisal) in der evidenzbasierten Medizin[3]. Evidenz ist in diesem Zusammenhang definitionsgemäß als Erkenntnis aus klinischen Studien und nicht als selbstevident oder plausibel zu verstehen. Entsprechend der Fehleranfälligkeit unterschiedlicher Studienkonzepte haben sich deshalb so genannte Evidenzhierarchien etabliert, die nicht als starre Einteilung sondern eher als „Daumenregel" zu verstehen sind [12]. Als Beispiel einer Evidenzskala sei diejenige des Centre for Evidence based Medicine in Oxford aufgeführt (siehe Tabelle 2).

Evidenzlevel halten fest, auf welchem methodischen Niveau die Erkenntnisse gewonnen wurden und wei-

[3] Nach David L. Sackett ist evidenzbasierte Medizin (EbM) definiert als der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen wissenschaftlichen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten. Die Praxis der EbM bedeutet die Integration individueller klinischer Expertisen mit der bestmöglichen externen Evidenz aus systematischer Forschung. Ein Zuwachs an Expertise spiegelt sich auch in der mitdenkenden und -fühlenden Identifikation und der Berücksichtigung der besonderen Situation, der Rechte und Präferenzen von Patienten bei der klinischen Entscheidungsfindung wider.

sen damit darauf hin, wie „robust" die gewonnenen Ergebnisse sind. Ihre Funktion ist vielfältig, sie dienen einerseits dazu, aus der Flut von Informationen eine Fokussierung auf diejenigen Studien vornehmen zu können, deren Aussagekraft am höchsten ist. Andererseits sind sie eine Hilfe, um sich widersprechende Erkenntnisse über eine Methode neu zu bewerten. Dies heißt nicht, dass die Ergebnisse von Studien niedriger Evidenzstufen den Ergebnissen von randomisierten, kontrollierten Studien immer widersprechen müssen, im Gegenteil ergeben sich häufig gleichgerichtete Ergebnisse [13, 14]. Dabei gilt jedoch, dass den Ergebnissen des methodisch höheren Niveaus im Zweifelsfall die weitergehende Erkenntniskraft zukommt.

Es wichtig festzuhalten, dass eine randomisierte kontrollierte Studie allerdings nur dann das höchste Studienniveau darstellt, wenn die Studiendurchführung entsprechend hochqualitativ war. So ist aus mehreren Untersuchungen bekannt, dass beispielsweise bei fehlerhafter Randomisation häufig eine Überschätzung des Therapieeffektes festzustellen ist [15, 16]. Auch wenn die Randomisation fehlerhaft verlaufen ist, so kommt der Studie noch zumindest das Erkenntnisniveau einer Kohortenstudie zu. Die einzelnen Therapiegruppen einer Studie können dann zwar nicht auf dem Niveau eines randomisierten, wohl aber eines kontemporären Vergleiches gegenübergestellt werden. Dies führte dazu, dass beispielsweise die Evidenzskala des Centre for Evidence Based Medicine in Oxford die explizite Zuweisung eines niedrigeren Evidenzniveaus für randomisierte Studien vorsieht, wenn bestimmte Qualitätskriterien nicht erfüllt sind (siehe Tabelle 2). Die Studienqualität bezieht sich ebenso auf die entsprechend angemessen formulierte Forschungsfrage, die dazugehörige Fallzahlberechnung, die Auswahl geeigneter Ergebnisparameter (Problem der sog. Surrogatparameter, s. auch [17, 18]) und die dazugehörige statistische Auswertung. Die Studienqualität nimmt selbstverständlich ebenso Einfluss auf alle anderen Studientypen. Auch bei Kohortenstudien ist beispielsweise der Zusammenhang zwischen fehlerhafter Durchführung und Ergebnisüberschätzung bekannt [19].

Es lässt sich festhalten, dass sich erst in der Gesamtschau von Studienkonzept (und dessen Angemessenheit zur Beantwortung einer klinischen Fragestellung), Studienqualität und klinischer Relevanz des Ergebnisses eine umfassende Beantwortung der eingangs formulierten drei Fragestellungen nach der Verlässlichkeit des Effektes einer medizinischen Intervention nach der Größe des beobachtenden Effektes und nach seiner klinischen Relevanz erzielen lässt. Das Evidenzniveau beleuchtet dabei nur einen, wenn auch wichtigen, Aspekt der Wertigkeit von Studienergebnissen.

Probleme klinischer Studien

Nicht jede Fragestellung ist geeignet, um durch eine randomisierte Studie geklärt zu werden. Klassische epidemiologische Fragestellungen zum Einfluss von nicht „zuweisbaren" Risikofaktoren (z. B. Nikotinabusus) oder langfristige Erkenntnisse zu Nebenwirkungen können durch andere, hierfür besser geeignete Studientypen wie beispielsweise die Kohortenstudie geklärt werden. Nicht immer ist eine vollständige Verblindung möglich, ebenso nicht ein Placebovergleich oder aus verständlichen Gründen auch der Vergleich mit einer nicht zu behandelnden Kontrollgruppe. Viele Fragestellungen werden als Probleme deklariert, die nicht in klinischen Studien zu klären sind, bei genauerer Betrachtung jedoch oft nur Detailfragen der Studienplanung betreffen, die in der Regel lösbar sind.

„Die randomisierte Studie ist die fairste und effektivste Methode des Erkenntnisgewinns."

Ebenso wird wiederholt die Frage der Randomisation in den Vordergrund gestellt. Kann es ethisch sein, per Zufall Therapien Patienten zuzuordnen, scheinbar im völligen Widerspruch zur ärztlichen Kunst, bei der der Arzt immer weiß, was am besten für seine Patienten ist? In einer systematischen Übersichtsarbeit zur Ethik randomisierter Studien stellten Edwards et al. 1998 folgende Postulate auf [20]: Klinische Studien sollten früh im Lebenszyklus einer neuen Behandlungsmethode einsetzen; dem Patienten muss eine vollständige Aufklärung über vergleichende Studien, die Behandlungsmethode und den Inhalt der klinischen Studie zukommen; Vergleichsgruppen können nur dann eingesetzt werden, wenn die zu vergleichenden Interventionen (Verum/Placebo-Alternativbehandlung etc.) zum Zeitpunkt des Studienbeginns als gleichwertig gelten müssen; analog der Deklaration von Helsinki sollte allen Patienten die beste Behandlung (best care) zukommen.

Die Autoren schließen, auch im Hinblick auf häufig geäußerte Bedenken, dass die randomisierte Studie die „fairste und effektivste" Methode des Erkenntnisgewinns (im Rahmen des Health Technology Assessment[4] [HTA]) sei. Unter Berücksichtigung der gängigen Kodizes zur klinischen Forschung ist deshalb festzuhalten, dass die nicht wissenschaftliche Generierung von Wissen als „unethisch" anzusehen ist [21].

Andere Probleme beziehen sich auf die Durchführbarkeit von randomisierten kontrollierten Studien (RCTs). Insbesondere in Deutschland wird eine Forschungsskepsis sowohl unter Ärzten als auch unter Patienten beklagt, die es fast unmöglich macht, entsprechende Studien durchzuführen [22]. Die in Studien optimierten Therapiebedingungen führen andererseits dazu, dass die Übertragbarkeit der Ergebnisse auf den klinischen Alltag unter Umständen nicht gegeben ist. Wenig bekannt ist jedoch auch, dass Patienten im Rahmen kontrollierter klinischer Studien in der Regel bessere Behandlungsergebnisse erfahren als in der täglichen Routineanwendung [20].

Sonderfall Diagnostik

Der Nachweis der Effektivität einer diagnostischen Methode erfordert eine Erweiterung gegenüber der Erkenntnismethodik, die zur Wirksamkeitsuntersuchung eines therapeutischen Verfahrens herangezogen wird. Die Tabelle 3 gibt die verschiedenen Ebenen des Erkenntnisgewinns zu bildgebenden Verfahren nach Thornbury [23] wieder. Es wird deutlich, dass nach der Sicherung von physikalisch-technischen Aspekten einer diagnostischen Methodik sowie

[4] Unter Technology Assessment wird gegenwärtig eine umfassende Bewertung neuer oder bereits auf dem Markt befindlicher Technologien hinsichtlich ihrer physikalischen, biologischen, auch im engeren Sinn medizinischen, ihrer sozialen und finanziellen Wirkungen im Rahmen einer strukturierten Analyse verstanden.

Tabelle 3
Ebenen des Erkenntnisgewinns aus Studien zu bildgebenden Verfahren (nach [23])

Fragestellung	Zielgrößen
Technische Effektivität	Bildauflösung, Bildschärfe, etc.
Diagnostische Genauigkeit	Sensitivität, Spezifität
Diagnostische Effektivität	Anteil der Fälle in einer Fallserie, bei denen die Diagnostik als hilfreich eingeschätzt wurde, Likelihood-ratios
Therapeutische Effektivität	Anteil der Fälle in einer Fallserie, bei denen die Diagnostik als hilfreich zur Therapieplanung eingeschätzt wurde bzw. zu Therapieänderungen führte
Outcome-Effektivität	Anteil der Patienten, die von der Anwendung der Diagnostik im Vergleich zu einer Nicht-Anwendung profitierte (kontrollierte Studie)

der Bestimmung der klassischen Parameter Spezifität und Sensitivität eine Prüfung des Verfahrens unter Berücksichtigung der therapeutischen Konsequenzen, die sich aus den Ergebnissen des diagnostischen Verfahrens für die Patienten ergeben, erforderlich ist. Provokant formuliert könnte gesagt werden, dass bei der Bewertung einer diagnostischen Methode in klinischen Studien in scheinbar paradoxer Umkehr des alten medizinischen Grundsatzes „vor der Therapie steht die Diagnose" nun „die Therapie vor der Diagnostik" steht. Der Nutzen eines diagnostischen Verfahrens hängt also wesentlich davon ab, ob der Patient von der Kenntnis des Ergebnisses eines diagnostischen Verfahrens in dem Sinne profitiert, dass wirksame Therapien eingeleitet werden bzw. unnötige Therapien oder weitere diagnostische Prozeduren vermieden werden können.

Idealerweise – aber bisher selten umgesetzt – ist die kontrollierte Studie auch zum Wirksamkeitsnachweis diagnostischer Verfahren als höchster methodischer Standard anzusehen. Es muss jedoch zwischen zwei Konzepten unterschieden werden. Einerseits sollte ein diagnostisches Verfahren in einer kontrollierten Therapiestudie als ein grundsätzliches Einschlusskriterium gedient haben. Als Beispiele können hier Studien zur Wirksamkeit von Bisphosphonaten bei Patienten mit Vorfrakturen und einer erniedrigten Knochendichte genannt werden [24]. Andererseits erbringen Studien, in denen bei einer Gruppe von Patienten das diagnostische Verfahren zu Therapieentscheidungen genutzt wird und in der Vergleichsgruppe lediglich die bisherigen diagnostischen Standards zur Anwendung kommen, einen noch weitergehenden Erkenntnisgewinn. Hier mögen jedoch sehr variable therapeutische Konsequenzen in den verschiedenen Diagnosegruppen zu ernsten Problemen in der Studiendurchführung und Studieninterpretation führen. Beispiele solcher Studien finden sich im Bereich der Prüfung der Effektivität der Resistenzbestimmung bei HIV-Patienten [25].

Fazit für die Praxis

Nicht alle Erkenntnisse zu medizinischen Methoden und Verfahren besitzen die gleiche Aussagekraft und Validität. Nach derzeitigen Vorstellungen stellt aus methodischen, nicht aus praktikablen Gründen, die randomisierte, kontrollierte Studie den Goldstandard für den Nachweis der Effektivität medizinischer Maßnahmen dar. Liegen zu einer Methode verschiedene Informationen und Ergebnisse vor, so kann der Evidenzhierarchie folgend eine Auflistung dieser Quellen erfolgen. Dabei kommt, sofern vorhanden, dem RCT die höchste Aussagekraft zu, gefolgt von anderen Studientypen (Fallserien, pro- oder retrospektiv angelegt) bis hin zur Expertenaussage. Diese Evidenzhierarchie kann eine Erleichterung bei der Bewältigung großer Informationsmengen, z. B. bei der Bewertung von Methoden im Bundesausschuss der Ärzte und Krankenkassen oder der Erstellung von HTA-Berichten, darstellen. Aus diesem Grund wurden Evidenzhierachien beispielsweise in der Verfahrensrichtlinie des Arbeitsausschusses „Ärztliche Behandlung" etabliert. Neben der Berücksichtigung der Wertigkeit des Studienkonzeptes (RCT, Kohortenstudien, Fallserie etc.) hat sich in der praktischen Anwendung gezeigt, dass alle Studien detailliert auf ihre ordnungsgemäße Durchführung und qualitative Wertigkeit geprüft werden müssen. Nur so kann vermieden werden, dass schlecht oder fehlerhaft durchgeführten Studien ein fälschlich hohes Evidenzniveau zugeschrieben wird.

Literatur

1. Unschuld PU (2000) Antike chinesische Medizin: Die Vielfalt der Denkstile. Dtsch Z Akup 43:21–32
2. Martini P (1953) Die klinisch-therapeutische Forschung und das Experiment. In: Martini P (Hrsg) Methodenlehre der therapeutisch-klinischen Forschung. Springer, Berlin Göttingen Heidelberg, S 6–16
3. Kolmogorov AN (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin Göttingen Heidelberg
4. Popper KR (1994) Logik der Forschung. Mohr, Tübingen
5. National Health and Medical Research Council (2000) How to use the evidence: assessment and application of scientific evidence. 1–83.2000. Canberra, National Health and Medical Research Council. Handbook series on preparing clinical practice guidelines. Commonwealth of Australia
6. Mosely JB, Wray NP, Kuykendall D, Willis K, Landon G (1996) Arthroscopic treatment of osteoarthritis of the knee: a prospective, randomized, placebo-controlled trial. Results of a pilot study. Am J Sports Med 24:28–34
7. Annas GJ (1992) The changing landscape of human experimentation: Nuremberg, Helsinki and beyond. Health Matrix 2:119–140
8. Howard-Jones N (1982) Human experimentation in historical and ethical perspectives. Soc Sci Med 16 (1429):1448
9. Armitage P (1992) Bradford Hill and the randomised controlled trial. Pharmaceut Med 6: 23–27
10. Silverman WA (1998) Where's the evidence? Debates in modern medicine. Oxford University Press, Oxford New York Tokyo
11. Torgerson DJ, Roberts C (1999) Randomization methods: concealment. Understanding randomised controlled trials. BMJ 319:375–376

12. NHS Centre for Reviews and Dissemination (2001) Undertaking systematic reviews of research on effectiveness. In : Khan KS, ter Riet G, Glanville J, Sowden JA, Kleijnen J (eds) 4th report, 2nd edn. University of York, York, CRD Reports
13. Benson K, Hartz AJ (2000) A comparison of observational studies and randomized, controlled trials. NEJM 342:1878–1886
14. Kunz R, Oxman AD (2000) The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 317:1185–1190
15. Colditz GA, Miller JN, Mosteller F (1989) How study design affects outcomes in comparisons of therapies. I. Medical. Stat Med 8: 441–454
16. Moher D, Jadad AR, Tugwell P (1996) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352:609–613
17. CAPS (the cardiac arrhythmia Pilot Study Investigators) (1988) Effects of encainide, flecainide, imipramine and moricizine on ventricular arrhythmias during the year after acute myocardial infarction: the CAPS. Am J Cardiol 61:501–509
18. CAST (Cardiac Arrhythmia Supression Trial) (1989) Preliminary report: effect of Encainide and Flecainide on mortality in a randomized trial of arrythmia suppression after myocardial infarction. NEJM 321:406–412
19. Berlin JA, Colditz GA (1990) A meta-analysis of physical activity in the prevention of coronary heart disease. Am J Epidemiol 132:612–628
20. Edwards S, Lilford R, Braunholtz D, Jackson JC, Hewison J, Thornton J (1998) Ethical issues in the design and conduct of randomised controlled trials. Health Technol Assess 2
21. Köbberling J (1997) Der Wissenschaft verpflichtet – Eröffnungsvortrag zum 103. Kongress der Deutschen Gesellschaft für Innere Medizin. Med Klinik 92:181–189
22. EB (2001) Lungenkrebs: Klinische Studien dringend nötig. Dtsch Ärztebl 98:A1036
23. Thornbury JR (1994) Clinical efficacy of diagnostic imaging: love it or leave it. AJR 162:1–8
24. Black DM, Cummings SR, Karpf DB et al. (1996) Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Fracture Intervention Trial Research Group. Lancet 348:1535–1541
25. Durant J, Clevenbergh P, Halfon P et al. (1999) Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. Lancet 353:2195–2199
26. Chaoui R, Kunz R (2000) EbM in der Geburtshilfe Beispiel „Antibiotika bei frühem vorzeitigen Blasensprung". In: Kunz R, Ollenschläger G, Raspe H, Jonitz G, Kolkmann FW (Hrsg) Lehrbuch evidenzbasierte Medizin in Klinik und Praxis. Deutscher Ärzteverlag, Köln, S 234–240
27. Sackett D, Rosenberg W, Gray J, Haynes RB (1996) Evidence-based medicine: what it is and what it isn't. BMJ 312:71–72
28. Schwartz FW, Dörning H (1992) Evaluation von Gesundheitsleistungen. In: Andersen H, Henke KD, Schulenburg JMGvdH (Hrsg) Basiswissen Gesundheitsökonomie, Bd. 1. edition sigma, Berlin, S 175–200