Übertragung von Studienergebnissen auf den Versorgungsalltag – Die therapeutische Praxis profitiert von der methodischen Vielfalt randomisierter und nicht-randomisierter Forschungsansätze*

Karl Wegscheider

Klinische und epidemiologische Studien sind heute Standardinstrumente der Forschung im Gesundheitswesen. Die Situation ist gekennzeichnet durch ein kaum noch zu bewältigendes Überangebot an Studienergebnissen bei gleichzeitig weiter bestehenden großen Forschungslücken in nicht wenigen medizinischen Bereichen. Parallel zur quantitativen Ausbreitung hat sich die Methodik aber in vieler Hinsicht auch qualitativ verbessert. Dank großer Fortschritte in Statistik und Epidemiologie wie z. B. der Bewältigung von Longitudinalanalysen, Überlebenszeitanalysen, Clusterrandomisationen und Adaptiven Designs können Studien heute effektiver und besser an die Fragestellung angepasst durchgeführt werden. Im Ergebnis hat der Forscher die Wahl zwischen einer Vielzahl von Studientypen und Forschungsansätzen. Es scheint für jedes Gesundheitsproblem eine passende Studienlösung zu geben.

Die aktuellen Studienansätze sind teilweise experimentell oder hochspezialisiert, einige auch umstritten. Ein großer Teil der neueren Studienansätze, bei weitem aber nicht alle, lassen sich als Unterformen des inzwischen allgemein anerkannten Goldstandards der Evidenzklasse 1 verstehen, der klassischen randomisierten kontrollierten Studien (RCT) bzw. der darauf aufbauenden Meta-Analysen und Systematischen Reviews.

Erstaunlicherweise sind Arzt und Patienten dennoch nicht zufrieden. Einerseits können sie das Überangebot an Studienergebnissen kaum bewältigen, andererseits fühlen sie sich häufig in den Studien nicht gebührend repräsentiert. Sie haben beide das Gefühl, dass ihre persönliche Wirklichkeit in den Studien nicht gut widergespiegelt wird. Und sie fühlen sich oft mit ihren Problemen allein gelassen. Immer da, wo man gerade eine konkrete Frage hat, scheint sich eine Lücke in der Evidenz aufzutun. Die klinische Wirklichkeit, der man sich aktuell ausgesetzt sieht, scheint sich in den Studien nicht so recht abzubilden. Woher rührt diese verbreitete häufig unterschwellige Unzufriedenheit?

Eine dazu häufig geäußerte Erklärung verweist darauf, dass sich in der beschriebenen Studienvielfalt eigentlich viel zu viel Minderwertiges verberge, das zur Verwirrung beitrage. Man solle sich deshalb auf die erwiesenermaßen besten Studien beschränken, d.h. auf die klassischen RCTs und von diesen Studien dann nicht viele kleine, sondern wenige große, besonders saubere machen bzw. zur Kenntnis nehmen – und den Rest einfach vergessen; davon hätten auch Arzt und Patient am meisten.

Im Folgenden soll diesem populären Erklärungsversuch die These gegenübergestellt (und begründet) werden, dass die Beschränkung auf Goldstandardstudien die Lücke zwischen Forschung und Praxis in der Medizin eher vergrößern würde. Stattdessen – so die Gegenposition – sollten wir uns wieder stärker um die Pflege der nicht-randomisierten Studien kümmern, die wir in den letzten Jahren vernachlässigt haben. Das Ziel sollte nicht die Standardisierung und Begrenzung von Studientypen, sondern die flexible Weiterentwicklung und sinnvolle Nutzung neuerer Ansätze sein, damit auch Fragestellungen bearbeitet werden können, die sich im Rahmen von RCTs nicht beantworten lassen.

Der Beitrag ist wie folgt gegliedert: Zunächst wird beschrieben, wie wir,

* (nach einem Vortrag, gehalten auf dem 2. Wissenschaftliches Diskussionsforum zur Nutzenbewertung im Gesundheitswesen des Gesundheitsforschungsrats (GFR) und des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) im Oktober 2008, [1])
angeforderter Artikel

DOI 10.3238/zfa.2009.0053

unabhängig vom Studientyp, prinzipiell aus klinischen oder epidemiologischen Studien lernen. In diesem Zusammenhang werden insbesondere die Problematik der Übertragbarkeit von Studienergebnissen auf die therapeutische Praxis sowie die Fehleranfälligkeit von Studien erörtert. Hierzu ist es hilfreich, die Natur des Forschungsprozesses grundsätzlich zu reflektieren.

Anschließend wird das schwierige Verhältnis von Wissenschaft und gesellschaftlicher Regulation erörtert und an einem Beispiel illustriert.

Danach wird zusammengetragen, welche Information und welche Studientypen Patient und Arzt für ihre therapeutische Entscheidungsfindung tatsächlich benötigen. Dem folgt die Untersuchung, ob Arzt und Patient in der aktuellen Forschungslandschaft tatsächlich mit dem versorgt werden, was sie brauchen. Aus all den Überlegungen werden schließlich allgemeine Empfehlungen zur künftigen Ausrichtung der Forschungslandschaft im Gesundheitswesen abgeleitet.

Wie wir aus klinischen Studien (nicht) lernen können

Das Kapitel „Klinische Studien" nimmt in der im Umfang eher begrenzten Methodikausbildung von Medizinstudenten und Prüfärzten einen besonderen Platz ein. Nach dem Kennenlernen der Vorteile sollte in einer guten Ausbildung auch über die Grenzen gesprochen werden, die vermittelten Konzepten eigen sind. In der Hamburger Biometrieausbildung werden hierzu fünf Begrenzungen diskutiert, die klinischen Studien unabhängig vom Studientyp gesetzt sind. Es handelt sich dabei ausnahmslos um Übertragungsprobleme im weiteren Sinne.

1. Aus Studien folgen statistische Aussagen (Kollektiv-Aussagen), keine individuellen Aussagen.

Von Florence Nightingale stammt der Satz: „Wenn wir wissen wollen, was Gott mit uns vorhat, müssen wir Statistik lernen."[1] Dieser Satz kann leicht missverstanden werden. Wenn überhaupt, dann erfahren wir durch Statistik, was Gott mit der Menschheit vor hat (darüber werden wir wohl verschiedene Meinungen haben), aber wir erfahren nichts über ein einzelnes Individuum (und das ist vielleicht auch gut so, wenn es z. B. um den Zeitpunkt des Todes geht). Patient und behandelnder Arzt befinden sich aber in einer Situation, in der es hauptsächlich um das Individuum gehen sollte. Leider können aber Studienergebnisse nicht so ohne weiteres auf ein konkretes Individuum übertragen werden.

2. Statistische Aussagen sind von Natur aus unscharf.

Wie immer wir es anstellen, die quantitativen Ergebnisse einer klinischen Studie werden unvermeidbar mit einer Unsicherheit belegt sein, die sich aus der Stichprobenziehung ergibt. Diese Unsicherheit kann beträchtlich sein und bis zur Nutzlosigkeit von Aussagen führen. Konfidenzbereiche sind ein Versuch, sich diese Unsicherheit vor Augen zu führen. Das kann sehr desillusionierend sein. Wir gucken da oft nicht so gern hin und ziehen es daher vor, p-Werte zu verwenden, die uns die Unsicherheit von Aussagen weniger aufdrängen und uns anschließend einen sorgloseren Umgang mit den Mittelwerten ermöglichen.

Mittelwerte sind aber ohnehin nur ein Teil der Wahrheit. Interessant ist die ganze Bandbreite biologischer Reaktionen auf eine Therapie, ausgedrückt in der gesamten Verteilung oder zumindest der Varianz. Varianzen sind aber noch schwieriger zu schätzen als Mittelwerte. Meist sind unsere Studien daher für viele klinisch wichtige Fragen nicht trennscharf genug.

3. Studienaussagen gelten nicht notwendig für andere Settings. (Externe Validität)

Dieser Gedanke ist inzwischen vielen vertraut. Die meisten Autoren denken an die externe Validität, wenn sie von Übertragbarkeitsproblemen sprechen. Das, was an einer Krankenhauspopulation von Patienten mit Herzinfarkt in einer Studie belegt ist, muss nicht für die Herzinfarktpatienten in der Hausarztpraxis gelten. Und das, was für solche gilt, die keine zusätzliche Lungenerkrankung haben, muss für die mit gleichzeitiger Lungenerkrankung nicht unbedingt auch gelten. Tatsächlich ist jedoch die externe Validität nur eines von mehreren Übertragungsproblemen, die wir bei klinischen Studien haben.

4. Studienaussagen gelten nicht notwendig für alle Untergruppen.

Ein besonders tückisches Problem ist die möglicherweise fehlende Übertragbarkeit nach innen. Wir sehen in unseren Studien die Patientenkollektive häufig als homogene Einheiten an. Tatsächlich ist jedoch keineswegs gewährleistet, dass die Gesamtaussage der Studien auch für alle Untergruppen gilt. Inhomogenitäten (Subgruppeneffekte) sind sogar zu erwarten. Meist gelingt es in einer Studie jedoch nicht, abweichende Subgruppen zu identifizieren, da der Stichprobenumfang nur für die Analyse der Gesamtgruppe berechnet wurde. Wir müssen also – häufig wider besseres Wissen – so tun, als ob ein Studienergebnis für z. B. Patienten mit Colon irritabile für alle Patienten gilt, die diese Diagnose haben.

5. Studienaussagen müssen nicht ewig gelten.

Tatsächlich ist auch die Übertragung von Studienergebnissen von einer Periode auf eine andere äußerst problematisch. Die Bevölkerung, die Lebensumstände sowie das medizinische Umfeld entwickeln sich. Wenn eine Therapie gestern hochwirksam war, muss dies heute nicht mehr gelten. Ein solcher schleichender Wirksamkeitsverlust wird „Creeping" genannt. Die FDA (Food and Drug Administration, US-Gesundheitsbehörde) hat diesen Effekt inzwischen entdeckt. Sie und andere Gesundheitsbehörden verlangen bei den sog. Nichtunterlegenheitsstudien daher einen direkten Nachweis, dass das gewählte Vergleichspräparat noch als wirksam anzusehen ist. Andererseits verlangen wir keine Neuzulassungsprozedur, wenn das

[1] Englische Übersetzung: To understand God, we must study statistics, for these are the measure of his purpose."

Altpräparat den Wirksamkeitsnachweis nicht schafft. Das führt zu dem paradoxen Effekt, dass alte Präparate trotz Zweifeln an der Wirksamkeit auf dem Markt bleiben und nicht einmal durch neuere äquivalente Präparate mit weniger Nebenwirkungen oder leichterer Anwendbarkeit ersetzt werden können.

In der Summe ist festzuhalten, dass es nicht nur ein Übertragungsproblem gibt, sondern viele Probleme mehr, die deutlich machen, dass ein Studienergebnis für einen sehr kleinen Ausschnitt einer Fragestellung nur, und dann auch nur als einigermaßen sicher gelten kann. Diese Probleme haben zudem gemeinsam, dass es keine empirische Methode gibt, um festzustellen, ob die Übertragbarkeit auf den zu versorgenden Patienten oder eine andere Patientengruppe gerechtfertigt oder zweifelhaft ist. Um herauszufinden, ob ein Studienergebnis auf ein anderes Studiensetting, eine Untergruppe, eine zukünftige Kohorte übertragbar ist, müsste man eine ausreichend große neue Studie in diesem Setting, dieser Untergruppe und dieser Kohorte durchführen. Dann müsste man aber die Übertragbarkeit ja gar nicht mehr beurteilen, weil es ja damit dann direkte Beobachtungen aus dem Zielkollektiv gäbe. Es folgt, dass die Übertragbarkeit von Studienergebnissen grundsätzlich nicht empirisch prüfbar ist.

Es sei im Übrigen darauf hingewiesen, dass dieses Problem den Pionieren der modernen klinischen Forschung durchaus bereits in seiner ganzen Problematik vertraut war. Ein Versuch, mit diesem Problem pragmatisch umzugehen, stellt das Prinzip der doppelten Wiederholung dar, dass seit langer Zeit als eine der Richtlinien bei der Bewertung der Evidenzlage in der Arzneimittelzulassung Verwendung findet: Ergebnisse müssen sowohl auf Patientenebene reproduziert werden (ausreichende Fallzahl) als auch auf Studienebene (in der Regel mindestens zwei randomisierte Studien in verschiedenen Settings).

Fehleranfälligkeit

Eine weitere wichtige Begrenzung ergibt sich daraus, dass es in der Praxis außerordentlich schwierig ist, Studien fehlerfrei durchzuführen. In Zulassungsverfahren gibt es praktisch keine Studien, die ohne kritische Relativierungen der Methodiker durch den Review kommen. Das gilt auch für randomisierte Studien. In den Hamburger Prüfarztkursen wird den Studierenden eine Übersicht über 12 Schwachstellen gegeben, an denen auch randomisierte Studien fundamental scheitern können und schon gescheitert sind. Fehlende Randomisation oder Fehler bei der Randomisierung bilden lediglich eine der Schwachstellen. Alle anderen möglichen Fehlerquellen wie Bias durch fehlende Verblindung, unterschiedliche Follow-up-Raten (differential missings), fehlerhafte Ermittlung der Zielgröße, übersehene Heterogenität und viele andere mehr betreffen randomisierte wie nicht-randomisierte Studien gleichermaßen Die Randomisation sorgt lediglich für die Vergleichbarkeit der Ausgangsbedingungen zu Beginn der Studie, schützt aber nicht vor anderen Schwachstellen.

Der Wissenschaftsprozess

Warum machen wir, wenn das alles denn so schwierig ist, überhaupt noch Studien? Auf diese nahe liegende Frage gibt es eine defensive und eine offensive Antwort:

1) Studien sind bei einer Vielzahl von medizinischen Problemen unsere einzige Chance, zu lernen und besser zu werden.
2) Es funktioniert! Studien geben unserem klinischen Verständnis wertvolle Impulse, obwohl sie selten klare und einfache Schlüsse zulassen. Sie „erden“ unsere Abbilder, die wir im Kopf tragen. – Dafür ist allerdings eine saubere Methodik unerlässlich.

Dass wir trotz der genannten Generalisierbarkeitsprobleme und trotz der grundsätzlichen Zweifel an der Aussagekraft aus Studien viel lernen können, verdanken wir den Besonderheiten des Wissenschaftsprozesses.

Das Fortschreiten des wissenschaftlichen Erkenntnisprozesses lässt sich gut durch eine Spirale veranschaulichen, die nach unten, d. h. in Richtung der Erkenntnis, immer breiter wird (Abb. 1). Die Bewegungsrichtung verläuft hauptsächlich in Schleifen. Die Wissenschaft kreist für eine längere Zeit beständig um die gleichen Fragen. Mal dominiert die eine, mal die andere Auffassung im ununterbrochen heftig geführten Diskurs, ein Markt der Meinungen und Einschätzungen, der von keinem wirklich beherrscht wird, obwohl es viele versuchen. Kurzfristig betrachtet scheint sich die Diskussion nur in die Breite zu bewegen. Längerfristig bewegt sich der Prozess aber in Richtung größerer Erkenntnis, in die immer mehr Gesichtspunkte einbezogen werden. Dabei werden zu keinem Zeitpunkt endgültige Entscheidungen getroffen. Alles ist vielmehr vorläufig und kann jederzeit relativiert werden. Strittige Fragestellungen werden jedoch im Laufe der Zeit wieder verlassen. Eine bestimmte Auffassung setzt sich dann durch, ohne jemals offiziell zur Wahrheit erklärt worden zu sein, und wird zur Grundlage weiterer Überlegungen. Eigentlich wird dabei nie eine Frage wirklich entschieden. Diskurse gehen häufig erst dadurch zu Ende, dass eine neue Generation von Wissenschaftlern die restliche Uneinigkeit nicht mehr so wichtig findet.

Der Prozess wird genährt durch ständige Impulse einer Vielzahl von empirischen Studien unterschiedlicher Qualität. Dabei wechseln sich Hypothesengenerierung (Exploration) und Hypothesenvalidierung (Konfirmation) in schneller Folge ab. In der Exploration liegt die Kreativität, in der Konfirmation die Anbindung an die Realität. Wissenslücken sind dabei ein normaler Zustand und werden durch Hypothesenbildung oder Analogieschlüsse gefüllt. Die Fähigkeit zu empirisch nicht oder noch nicht abgesicherten Analogieschlüssen ist dabei das eigentliche Charakteristikum und eine Stärke, nicht eine Schwäche der Wissenschaft, ein Charakteristikum, das als solches nur dem Menschen eigen ist. Im Zusammenwirken mit dem kritischen Geist wird der Mensch durch diese Fähigkeit in die Lage versetzt, Gesetzmäßigkeiten und ihre Grenzen zu erkennen. Der Umstand, dass das grundsätzlich nur unscharf und fehlerbehaftet möglich ist, ändert nichts an der beispiellosen Erfolgsgeschichte von Wissenschaft, ohne die menschliches Überleben in der heutigen Form kaum denkbar wäre.

Eine Erfolgsgeschichte ist auch die Entwicklung der Instrumente ‚Klinische Studien‘ und ‚Klinische Epidemiologie‘, ohne die ein Großteil der heutigen medizinischen Praxis nicht in der Form existieren würde, die wir inzwischen als selbstverständlich hinnehmen. Dabei wird allerdings, solange

Abbildung 1 Darstellung des Wissenschaftsprozesses und seiner Impulsgeber.

man sich auf RCTs beschränkt, die therapeutische Wirklichkeit nur sehr ungleichmäßig ausgeleuchtet. Neben therapeutischen Fragestellungen, auf die große randomisierte Studien ein scharfes Schlaglicht richten, finden sich teilweise in unmittelbarer Nachbarschaft alltägliche Fragestellungen, für die es keine Studien der Evidenzklasse 1a gibt.

Nicht-randomisierte Studien können dieses Dunkel zwar nicht hell ausleuchten, sie schaffen aber weitere schwächer erleuchtete Stellen. Die restlichen Lücken müssen vom behandelnden Arzt an und mit seinem Patienten ausgefüllt werden. Er füllt sie ebenfalls mit Analogieschlüssen, das heißt er unterstellt Übertragbarkeit auf der Basis von Plausibilitätsüberlegungen. Dies ist zwar nicht im engeren Sinne evidenzbasiert, aber dennoch eine wissenschaftliche Vorgehensweise. Damit liegt dieser Bereich zwar zweifellos eher in einem Dämmerlicht, und wir haben Mühe, klar zu erkennen, was sich dort verbirgt. Dieser Zustand ist aber immer noch besser als die völlige Dunkelheit, die an diesen Stellen zu finden wäre, wenn wir uns nicht-randomisierte Studien und Analogschlüsse verbieten würden.

Wissenschaft und gesellschaftliche Regulation

Mit dieser Situation können die meisten Akteure gut leben. Ein Problem mit dieser Situation haben hingegen die Regulatoren, die sich zum Ziel gesetzt haben, Therapiebewertungen hauptsächlich auf der Basis hochrangiger Evidenz vorzunehmen, sei es bei der Zulassung oder bei der Kostenerstattung. Für sie ist der freie Analogieschluss unzureichend, weil er schwer objektivierbar und interpretationsabhängig ist. Man kann sich zunächst das Problem eine Zeitlang vom Leib halten, indem man den Antragstellern, die eine Zulassung oder Kostenerstattung anstreben, hohe Nachweishürden setzt. Auf diese Weise schafft man sich eine idealisierte Subwelt, in der man sich der Illusion der zumindest statistischen Kontrolle von Schlussfehlern hingeben kann. Spätestens mit der Ausweitung des Bewertungsauftrages über den Wirksamkeitsnachweis hinaus auf eine vergleichende Nutzenbewertung ist diese Position nicht mehr durchzuhalten. Die überall lauernden Evidenzlücken machen die Frage der Übertragbarkeit und den Wunsch nach einer wissenschaftlichen Methodik hierfür dringend. Der Regulator (z. B. der GBA) kann (und sollte vernünftigerweise) den Bewertungsauftrag an ein wissenschaftliches Institut verschieben (z. B. indem man das IQWiG mit einem Nutzenbericht beauftragt). Damit ist das Problem aber noch nicht gelöst.

Eine Folge der Konzentration auf Studien höchster Evidenzklassen (RCTs) und darauf aufbauende Meta-Analysen ist die Einengung des Blickwinkels auf stark zusammenfassende Maßzahlen (Mittelwerte, Signifikanzen, Meta-Mittelwerte, Meta-Signifikanzen). Damit kann aber auch eine vom Wesentlichen ablenkende Reduktion der Information verbunden sein.

Ein Beispiel: Die TORCH-Studie

Als Beispiel sei hier die TORCH-Studie angeführt [2]. In dieser Studie wurde eine fixe Kombination von zwei Asthma-Medikamenten mit unterschiedlichen Wirkpfaden (Salmeterol (S) und Fluticasone proprionate (F)) bei COPD-Patienten gegen Placebo und gegen die einzelnen Komponenten getestet. Hauptzielgröße war die Gesamt-Mortalität. Der Sponsor strebte die Zulassung der Kombination zum Zweck der Mortalitätsreduktion an. Entsprechend wurde die Studie auf den Vergleich zu Placebo optimiert. Hierzu wurde ein gruppensequentielles Design verwendet mit dem Ziel, die erforderliche Patientenanzahl zu minimieren. Im Ergebnis wurde die formale Signifikanz knapp verfehlt. Der Autor des zugehörigen Editorials [3] vermerkt: "In the end, the trial failed to meet its goal: the p value for death from any cause was 0.052, which was higher than the pre-specified value of 0.050. All clinical trials are a gamble, and the TORCH investigators came close to winning but did not win. Thus, the results of this trial are difficult to interpret."

Mit diesen Sätzen spiegelt der Editor die Blickverengung der Primäranalyse. Das Ergebnis der klinischen Studie wird auf einen p-Wert reduziert. Die Frage, ob der p-Wert knapp oberhalb oder unterhalb von 0,05 liegt, wird zum dominierenden Gegenstand der Betrachtung. Die klinische Studie wird zum Spielball der Zocker.

Die wirkliche Bedeutung der TORCH-Studie liegt hingegen auf einer anderen Ebene. Als erste große randomisierte Mortalitätsstudie in ihrem klinischen Setting gab sie der wissenschaftlichen Diskussion über die richtige Behandlung von COPD-Patienten einen kräftigen Impuls. Während man vorher vermutet hatte, dass die Mortalitätsreduktion eher durch die Kortikosteroide (F) als durch langwirkende Beta-Agonisten (S) bewirkt wird, ließ die Studie erkennen, dass tatsächlich nur Beta-

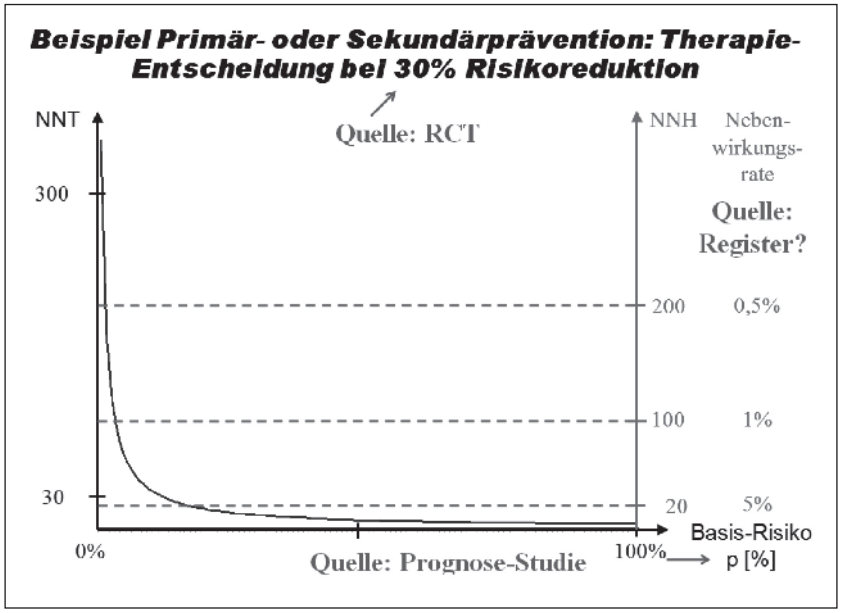

Abbildung 2 Schematische Darstellung des therapeutischen Entscheidungsprozesses bei der Prävention: Ermittlung des individuellen Nutzens durch die Number needed to treat (NNT) und des potentiellen Schadens durch die Number needed to harm (NNH) für einen Patienten mit Basisrisiko p.

Agonisten die Sterblichkeit reduzieren. Beide Substanzen wirken hingegen symptomlindernd. Die Medikamente interagieren nicht und können frei kombiniert werden. Diese für die Behandlungspraxis wichtigen Aspekte ließen sich jedoch aus der Studie nicht sicher schlussfolgern, da die Power dafür nicht ausreichte. Vielmehr müssen supportiv weitere Studien hinzutreten, die diese Schlüsse erhärten oder relativieren können. Der tatsächliche Wert der Studie liegt in der Diskussion dieser Aspekte durch die Fachwelt, nicht in der evidenzbasierten Beurteilung der Fixkombination – oder dem Umstand, dass der gewünschte p-Wert gerade nicht erreicht wurde.

Welche Informationen brauchen Arzt und Patient aus klinischen Studien?

Individueller Nutzen und Sicherheit des Patienten

Mit den Evidenzlücken hat nun aber nicht nur der Regulator, sondern insbesondere auch der Arzt und sein Patient Probleme. Sie stehen wie der Regulator und anders als die Wissenschaft unter Entscheidungszwang. In gewisser Weise ist ihre Situation noch viel dramatischer als die des Regulators: während dieser zur Not, wegen schlechter Evidenzlage, auch Zuflucht zu einer pragmatischen politischen Entscheidung nehmen könnte, müssen Arzt und Patient eine Therapieentscheidung (die auch Zuwarten sein kann) treffen, die weit reichende Konsequenzen haben kann.

Und der behandelnde Arzt hat zudem eine weitaus schwerere Aufgabe zu lösen als der Wissenschaftler oder der Regulator. Er muss sich nicht nur Gewissheit über die Wirksamkeit einer Therapie verschaffen, sondern darüber hinaus einschätzen, ob die Therapie seinem individuellen Patienten nutzen wird und ob sie ihm schaden kann. Dabei geht es anders als in den IQWiG-Nutzenberichten nicht um einen kollektiven Nutzen, sondern um einen ganz persönlichen Nutzen. Die entscheidende Schlüsselzahl hierfür ist nicht die Risikoreduktion, die wir zurecht zur Beurteilung der Wirksamkeit heranziehen, sondern die sog. „number needed to treat (NNT)“, die die Zahl der Patienten angibt, die man behandeln muss, damit einer von ihnen profitiert. Die NNT hängt entscheidend vom Basis-Risiko des Patienten (dem Risiko ohne Behandlung) ab, muss also als Funktion dieses Basisrisikos abgetragen werden, weil sie für jeden Patienten anders ist.

Diese Kurve hat einen bemerkenswerten Verlauf. Bei großem Risiko ist die individuelle NNT so niedrig, dass eine Behandlung sehr empfehlenswert erscheint. Bei sehr kleinem Ausgangsrisiko (z. B. bei einem jungen sonst gesunden Patienten mit nur einem Risikofaktor) erscheint eine Behandlung bei sehr hohen NNT-Werten unwirtschaftlich.

In dem Bereich dazwischen aber hängt die NNT extrem vom genauen Zahlenwert des Ausgangsrisikos ab. Für die Behandlungsentscheidung ist in diesem Bereich, in dem häufig ein großer Teil der Patienten, insbesondere der ambulanten zu finden ist, die Beurteilung des Basis-Risikos des Patienten wichtiger als die genaue Bezifferung der Risikoreduktion.

Bei der Behandlungsentscheidung müssen Arzt und Patient darüber hinaus berücksichtigen, welche Nebenwirkungen unter der Therapie zu erwarten sind. Das geschieht am besten durch die zusätzliche Betrachtung der „number needed to harm (NNH)“, die häufig, aber nicht immer, unabhängig vom Basis-Risiko ist und dann als Grundlage für therapeutische Entscheidungen zur NNT in Relation gesetzt werden muss.

Abbildung 2 stellt dar, wie diese Abwägung funktioniert. Die Kurve stellt die Abhängigkeit der NNT vom Basisrisiko für eine Therapie dar, die laut Studienlage das zu behandelnde Risiko um 30 % reduziert. Die NNT liegt z. B. bei 50 %igem Basisrisiko unter 15. Bei einem Basisrisiko von z. B. nur 1 % liegt sie hingegen über 300.

Rechts ist jeweils die NNH drei unterschiedlich häufigen Nebenwirkungen (0,5 bis 5 %) zugeordnet, deren Auftretenshäufigkeiten vom Basisrisiko aber unabhängig sind (gestrichelte Linien). Am Schnittpunkt mit der NNT-Kurve gilt jeweils: pro verhindertem kritischem Ereignis muss mit ebenfalls einem Patienten mit einer entsprechenden Nebenwirkung gerechnet werden. Bei kleineren Basisrisiken überwiegt die Zahl der Nebenwirkungen die der gewünschten Wirkungen; bei größeren Basis-Risiken hingegen überwiegt die Zahl der verhinderten Krankheits-Ereignisse die Zahl der Nebenwirkungen. Bei z. B. einem Basisrisiko von 20 % steht die Nebenwirkung mit Nebenwirkungsrate 5 % gerade der ge-

	Frage	Aspekt	Studientyp
Patient/Arzt	Was wird helfen? Was wird, wenn ich das Risiko nicht eingehe? Welche Risiken gehe ich dafür ein?	Wirksamkeit natürlicher Verlauf Sicherheit	RCT Kohorten-Studie RCT, Kohorten-Studie, Register
Patient	Wie ist meine Krankheit einzuordnen? Wie wird es mir unter der Therapie gehen?	Krankheitsverständnis Lebensqualität	Qualitative Studie Kohorten-Studie, RCT

Abbildung 3 Fragen bzw. Aspekte, die in den therapeutischen Entscheidungsprozess eingehen und für die Beantwortung erforderliche Studientypen.

wünschten Wirkung gleichauf gegenüber. Bei einem Basisrisiko von 50 % überwiegt hingegen ganz eindeutig die gewünschte Wirkung einer Behandlung die mögliche Nebenwirkung (hier im Beispiel die mit 5 % Häufigkeit). Bei selteneren Nebenwirkungen (hier z. B. 0,5 %) ist die gewünschte Wirkung bei nennenswertem Basisrisiko immer deutlich häufiger; ist aber das Basisrisiko nur um die 2 %, dann ist die NNT um 200, die Nebenwirkung (mit 0,5 % Wahrscheinlichkeit) führt ebenfalls zu einer NNH von 200.

Welche Relation von NNH und NNT für den Patienten noch vertretbar ist, hängt von der Schwere der Ereignisse und Nebenwirkungen sowie den Präferenzen des Patienten ab. Um den Patienten gut beraten zu können, muss sich der Arzt diese Darstellung sicherlich nicht formal erstellen. Er wird aber dieselben Überlegungen auch ohne formale Berechnungen anstellen, wenn er intuitiv die erforderlichen Abwägungen mit dem Patienten zusammen vornimmt.

Welche Studientypen werden benötigt?

Woher erhält der Arzt die Angaben, die er für den beschriebenen Abwägungsprozess braucht? Die Risikoreduktion erfährt er aus RCTs und Meta-Analysen. Das Basis-Risiko hingegen stammt eher aus Prognosestudien, die überwiegend als nicht-randomisierte Kohorten-Studien organisiert sind (Beispiel: Framingham-Studie). Nebenwirkungsraten können aus den RCTs vor Zulassung meist nicht realitätsnah und ausreichend präzise geschätzt werden. Hierzu sind qualitativ hochwertige Registerdaten und Angaben aus Pharmakovigilanz-Datenbanken erforderlich. Der Patient will darüber hinaus häufig noch mehr wissen, z. B., wie es ihm unter der Therapie gehen wird. Fragen nach der Lebensqualität können am besten aus im Gefolge von Interventions- oder Kohortenstudien durchgeführten wiederholten Befragungen (Surveys) beantwortet werden. Insbesondere bei chronischen Erkrankungen hat der Patient zudem ein Bedürfnis, seine Krankheit in sein Leben einzuordnen. Für die Aufgabe der Krankheitsbewältigung sind häufig um Verständnis bemühte qualitative Studien hilfreicher als quantitative Studien. Damit wird deutlich, dass wir verschiedene methodische Zugänge brauchen, um in Bezug auf Krankheiten und Therapien alles das aus Daten zu lernen, was man aus Daten lernen kann.

Insgesamt ist somit festzuhalten, dass randomisierte Studien nur einen Teil der Informationen bereitstellen können, die Patient und Arzt für eine Therapieentscheidung benötigen. Andere Studientypen müssen hinzutreten und können für die individuelle Entscheidung sogar bedeutsamer sein als die Quantifizierung der Wirksamkeit durch eine RCT. Bei der reinen Wirksamkeitsbeurteilung sind diese Studien den RCTs in Bezug auf die interne Validität zwar unterlegen und werden deshalb in niedrigeren Evidenzklassen geführt. Sie tragen jedoch andere für Therapieentscheidungen unverzichtbare Informationen bei. Zu betonen ist in diesem Zusammenhang auch, dass die beschriebene individuelle Nutzenabwägung originäre Aufgabe des behandelnden Arztes ist und nicht durch eine kollektive Nutzenentscheidung einer zentralen Institution ersetzt werden kann. Es ist hingegen sehr wohl vorstellbar, dass bei guter Studienlage eine Bestimmung des kollektiven Nutzens entfallen kann, weil der individuelle Nutzen gut beurteilbar ist.

Welche Informationen bekommen Arzt und Patient aus klinischen Studien? – Der Bias in der Forschungslandschaft

Bekommen Patient und Arzt nun die Studien, die sie brauchen? Über diese Frage haben sich die Mitglieder des Institute of Medicine (IoM) – eines großen virtuellen Institutes, dessen Mitglieder Verantwortung tragende Wissenschaftler aus den USA sind – Gedanken gemacht. Sie untersuchten die Hintergründe für die spektakulären Warn- und Rücknahme-Aktionen aus Sicherheitsgründen, die wir in den letzten Jahren im Arzneimittelbereich erlebt haben. Sie konstatierten einen kräftigen Bias in der Forschungslandschaft. Während die Wirksamkeit eines neuen Medikamentes generalstabsmäßig mit aktiver Datensammlung und sorgfältig abgestimmten und finanziell gut ausgestatteten Studienprogrammen bereits vor der Markteinführung erfolgt und dann auf eine gut ausgestattete und präparierte Behörde mit Sanktionsmöglichkeiten trifft, wird die Sicherheit nach der Zulassung nur mit passiven Datensammlungen und unterfinanziert untersucht. Nebenwirkungen werden nur zufällig entdeckt. Die unterausgestattete Behörde hat kein abgestuftes Instrumentarium zur Erkundung von und zum differenzierten Umgang mit Risiken. Patient, Arzt und die Gesellschaft als Ganzes bekommen so ein lückenhaftes und verfälschtes Bild von der Nutzen-Risiko-Bilanz von Medikamenten. Entsprechend fordert das IoM die Neuregulierung des gesamten Bereiches [4]. Hierzu werden

Prof. Dr. Karl Wegscheider...

... ist seit 30 Jahren als statistischer Berater und Studienbiometriker in der klinischen und epidemiologischen Forschung tätig, insbesondere in der Kardiologie, der Allgemeinmedizin und der Versorgungsforschung. Seine Forschungsinteressen liegen im Bereich der clusterrandomisierten Studien und der Longitudinalanalysen. Heute leitet er das Institut für Medizinische Biometrie und Epidemiologie am Universitätsklinikum Hamburg-Eppendorf. Nach einer schweren Erkrankung an Poliomyelitis ist er als Tetraplegiker auch seit 20 Jahren Patient. Bei der Bewältigung des Alltags als Berufstätiger und Familienvater ist er vielfältig auf Assistenz angewiesen.

im Einzelnen folgende Maßnahmen empfohlen: gleichmäßige Verteilung der FDA-Gelder zwischen den Bereichen Wirksamkeit und Sicherheit, Autorisierung der FDA, Post-Marketing-Studien zu verordnen und Konsequenzen aus den Ergebnissen zu ziehen, Ex-ante-Registrierung aller Studien, Verbesserung der Risikokommunikation mit Ärzten und Patienten, Kennzeichnung neuer Medikamente mit einem schwarzen Dreieck und für die ersten zwei Jahre ein Direct-to-consumer-Werbeverbot.

Während das IoM in der ungeeigneten Regulation die Hauptursache für den Bias in der Forschungslandschaft sieht, weist der italienische Kardiologe Picano, ein Spezialist für bildgebende Verfahren, auf die unterschiedliche Gewichtung von Studien unter Idealbedingungen zur Untersuchung der Wirksamkeit und Studien im Feld zur Untersuchung der Nützlichkeit hin. Wissenschaftlich attraktiv und karrierefördernd sind nur ideale Studien. Sie schüren Hoffnungen und erhöhen die Chancen auf wirtschaftliche Verwertung. Patientennahe Studien zur Abbildung der Versorgungswirklichkeit werden entsprechend zu wenig durchgeführt. Gerade diese Studien werden aber für Therapie-Entscheidungen im Einzelfall dringend benötigt.

Empfehlungen

Welche Empfehlungen können aus den vorstehenden Überlegungen abgeleitet werden?

1. Sicherlich ist es wünschenswert, dass wir uns von der ausschließlichen Orientierung an Wirksamkeitsnachweisen lösen und Versorgungsaspekte stärker in Studien zur Geltung kommen lassen.
2. Insgesamt wäre es vorteilhaft, wir würden weniger ergebnisorientiert denken, Komplexität und wissenschaftliche Lernprozesse stärker würdigen und die Bedeutung der diskursiven Rezeption von Studienergebnissen erkennen.
3. Wir sollten weniger auf Signifikanz orientierte, sondern mehr pragmatische Studien durchführen. Randomisierung bleibt ein wesentliches Element, das die Vergleichbarkeit innerhalb von Studien verbessert. Dennoch ist es sinnvoll, eine RCT zu ergänzen, indem man sie in Beobachtungsstudien oder Register integriert, die ähnlich hohen Qualitätsstandards genügen sollten. Von hoffnungsvollen Ansätzen für derart integrierte Forschungslandschaften wird zunehmend berichtet.

Interessenskonflikte: Laut Autor keine vorhanden.

Korrespondenzadresse:

Prof. Dr. Karl Wegscheider
Universitätsklinikum Hamburg-Eppendorf
Zentrum für Experimentelle Medizin
Institut für Medizinische Biometrie und Epidemiologie
Martinistr. 52, 20246 Hamburg
E-Mail: k.wegscheider@uke.de

Literatur

1. http://www.iqwig.de/index.810.html
2. Calverley PMA, Anderson JA, Celli B, et al. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. N Engl J Med 2007;356:775–89.
3. Rabe, KF. Treating COPD – The TORCH Trial, P Values, and the Dodo.[Editorial]. N Engl J Med 2007;356:851–54.
4. Psaty, Bruce M.; Burke, Sheila P. Institute of Medicine on Drug Safety. N Engl J Med 2007;355:1753–55.