Oliver Senn

Wie lese ich Studien? EBM als «Tool» anstelle «Dogma»

Workshop-Bericht von der SFD Conference 29.–30. August 2013 in Bern

Evidenzbasierte Medizin (EBM) ist die Forderung nach der Berücksichtigung der «best research evidence» als integralem Bestandteil der medizinischen Entscheidungsfindung mit Einbezug der klinischen Erfahrung und der Präferenzen des Patienten. «Skills» bei der kritischen Beurteilung wissenschaftlicher Evidenz (critical appraisal) können helfen, bei der Fülle der Informationen die «Spreu vom Weizen» zu trennen.

Einleitung

Unser (elektronisches) Postfach wird täglich mit einer Fülle von «neuesten medizinischen Erkenntnissen» und «noch besser wirksamen» Therapien gefüttert. Im Praxisalltag sind wir als verantwortungsvolle Ärzte und Ärztinnen stetig bemüht, die besten Informationen zu Diagnose, Therapie, Prognose und Prävention in die Betreuung unserer Patienten einfliessen zu lassen. Unabdingbar für eine gute Patientenbetreuung sind unsere klinischen Fähigkeiten und Erfahrungen. Nebst der klinischen Expertise spielt auch der Austausch in Peer Groups (z.B. im Rahmen des Qualitätszirkels) eine wichtige Rolle, wenn es darum geht, die praktische Relevanz neuer Studiendaten zu bewerten. Evidenzbasierte Medizin (EBM) ist die Forderung nach der Berücksichtigung der «best research evidence» als integralem Bestandteil der medizinischen Entscheidungsfindung zusammen mit der klinischen Erfahrung und den Wertvorstellungen und Präferenzen des Patienten [1]. Eine kritische Beurteilung der Evidenz (critical appraisal) hinsichtlich Validität, Relevanz und Praktikabilität ist somit Voraussetzung und Bestandteil einer evidenzbasierten Medizin. Nachfolgend sollen einige Tipps und Tricks aufgezeigt werden, welche helfen sollen bei der Fülle von Informationen die «Spreu vom Weizen» zu trennen.

Die Frage der Kausalität

Versuchen sie, sich folgende Situation vorzustellen. Sie sitzen mit ihrem Hund im Wohnzimmer und ein Fremder nähert sich ihrem Haus. Der Hund fängt an zu bellen. Eine Frage, die man sich in dieser Situation stellen könnte, wäre folgende: Falls der Fremde nicht vor ihrem Haus aufgetaucht wäre, hätte der Hund trotzdem angefangen zu bellen? Attestiert man dem Tier gute Qualitäten als Wachhund, scheint die Schlussfolgerung klar und der Fremde wäre als Ursache für das Bellen des Hundes verantwortlich. Für Skeptiker wäre auch der Zufall eine plausible Erklärung und für Leute mit negativen Erfahrungen mit Hunden kämen wohl auch andere Eigenschaften des Hundes für das Bellen in Frage. Dieses Beispiel illustriert, dass der Zusammenhang eines beobachteten Phänomens nicht unbedingt kausal sein muss. Sir Austin Bradford Hill, ein englischer Statistiker und Epidemiologe, hat 9 Kriterien formuliert [2], welche bei einer postulierten Kausalität aus Beobachtungen kritisch überdacht werden sollten (Tabelle 1). *Insbesondere bei Beobachtungsstudien gilt, dass eine Studie allein selten genügend robuste Evidenz liefert, die eine Änderung der klinischen Praxis rechtfertigen würde.*

Wenn es im Praxisalltag um die Prognose eines (Risiko)-Verhaltens geht oder um den Nutzen einer Therapiemassnahme, müssen wir uns unmittelbar mit der Frage der Kausalität auseinandersetzen. Unterschiedliche Studientypen befassen sich mit dem wissenschaftlichen Nachweis kausaler Zusammenhänge. Die Vor- und Nachteile der häufigsten Studientypen sollen nachfolgend kurz vorgestellt werden, wobei für umfassende Ausführungen auf Lehrbücher der Epidemiologie verwiesen wird [3, 4].

Studientypen

Im Kontext quantitativer Forschung lassen sich grundsätzlich experimentelle Studien von nicht-experimentellen (beobachtenden) Studien unterscheiden. Der Goldstandard für den Nachweis einer kausalen Beeinflussung eines Krankheitsverlaufs durch eine Therapiemassnahme ist die randomisierte (experimentelle) Studie. Die Kohortenstudie und die Fall-Kontroll-Studie sind häufige Designs, welche insbesondere bei der Erforschung der Ätiologie und der Risikofaktoren von Krankheiten angewendet werden.

Tabelle 1
Bradford-Hill Kriterien für Kausalität [2].

Stärke (strength)	Insbesondere eine scheinbar «schwache Assoziation» besagt nicht, dass keine Kausalität vorhanden ist.
Folgerichtigkeit (consistency)	Gibt es übereinstimmende Beobachtungen in anderen Studien?
Spezifität (specificity)	Gibt es Hinweise, dass eine spezifische Population (z.B. Raucher, spezifische Berufstätigkeiten) an einer Erkrankung leidet?
Zeitlichkeit (temporality)	Der zeitliche Zusammenhang sollte so sein, dass die Wirkung nach der vermuteten Ursache eingetreten ist.
Biologischer Gradient (biological gradient)	Ist eine Dosis-Wirkungs-Beziehung vorhanden (z.B. Anzahl Packyears und Krebsrisiko)?
Plausibilität (plausibility)	Ein plausibler (biologischer) Mechanismus für das Ursache-Wirkungs-Prinzip ist hilfreich.
Stimmigkeit (coherence)	Gibt es übereinstimmende Daten zwischen epidemiologischen Resultaten und Ergebnissen aus Laboranalysen?
Experiment (experiments)	Gibt es experimentelle Evidenz für eine Ursache-Wirkungs-Beziehung?
Analogie (analogy)	Gibt es Evidenz, dass ähnliche Risikofaktoren ähnliche Effekte haben?

Randomisierte Studie
Die zufällige (random) Aufteilung einer klar definierten Population in üblicherweise zwei Gruppen (Intervention und Kontrolle) ist das Grundelement der randomisierten Studie. Der entscheidende Punkt bei der (erfolgreich durchgeführten) Randomisierung ist, dass die Gruppenzuteilung nicht vorausgesagt werden kann und damit zu Beginn des Experiments zwei absolut vergleichbare Populationen vorhanden sind. Zeigen sich unterschiedliche Krankheitsentwicklungen bzw. Therapieerfolge in den beiden Gruppen, so ist das einzige Unterscheidungsmerkmal zwischen den Gruppen die Art der durchgeführten Intervention (z. B. Verum vs. Plazebo). Die randomisierte Studie hat eine hohe interne Validität, jedoch häufig eine beschränkte Generalisierbarkeit für die tägliche Praxis (externe Validität), da eine Fülle von Ein- und Ausschlusskriterien angewendet werden [5]. *Eine starke Forschung in der Hausarztmedizin kann dazu beitragen, dass das spezifische Setting (ältere, multimorbide PatientInnen) auch adäquat in solchen Studien berücksichtigt wird.* Damit ist die hausärztliche Versorgungsforschung in der Lage, genau jene Evidenz für den Praxisalltag zu generieren, die sich aus den klinischen Studien (die meist von der pharmazeutischen Industrie durchgeführt werden) nicht ableiten lässt.
Zufallsergebnisse und Subgruppen-Analysen
Da Krankheiten oft sehr unterschiedliche Ausprägungen haben, ist es verlockend, in spezifischen Subgruppen nach Therapieeffekten zu suchen. Die Präsentation von Subgruppen-spezifischen Effekten ist daher in randomisierten Studien häufig anzutreffen. *Resultate basierend auf Subgruppenanalysen haben jedoch eine grosse Gefahr, dass die beobachteten Effekte zufälliger Natur sind,* dies unabhängig von der Grösse der ursprünglichen Studienpopulation. Ein schönes Beispiel für die Zufälligkeit von Subgruppen Effekten zeigte die «International Study of Infarct Survival» (ISIS-2) [6]. ISIS-2 demonstrierte eindrücklich die Überlegenheit von Aspirin und Streptokinase auf das Überleben nach akutem Myokardinfarkt im Vergleich zu Plazebo bei insgesamt 17 187 randomisierten Patienten. Um auf die Trugschlüsse von Subgruppen Analysen aufmerksam zu machen, haben die ISIS-2 Autoren die Patienten nach ihren astrologischen Sternzeichen klassifiziert. Selbst bei der riesigen Fallzahl zeigte sich, dass in der Patientenpopulation, welche unter den astrologischen Sternzeichen «Zwillinge» oder «Waage» geboren wurden, Aspirin keinen Nutzen zeigte und sogar schädlicher war im Vergleich zu Plazebo. Für Nicht-Astrologen ist dieser Effekt sicher zufällig.
Ein Grund für diese zufälligen Effekte ist statistischer Natur, da es einen Zusammenhang gibt zwischen der Anzahl getesteter Hypothesen und der Wahrscheinlichkeit, ein signifikantes Resultat (P-Wert <0.05) zu bekommen (Abb. 1). *Resultate aus Subgruppenanalysen können helfen, (rationale) Hypothesen zu generieren, welche jedoch prospektiv an einer neuen Stichprobe getestet werden sollten, bevor kausale Schlussfolgerungen gemacht werden.* Bei der kritischen Beurteilung von Subgruppenresultaten sollte auf folgende Punkte geachtet werden:
– Gibt es einen (signifikanten) Effekt in der Gesamtpopulation?
– Wurden die geplanten Subgruppenanalysen «à priori» geplant?
– Gibt es eine plausible Erklärung/Hypothese für die gewählten Subgruppen?
– Wurden statistische Vorkehrungen getroffen (p-Wert Korrektur, Interaktions-Analysen) um das Risiko für zufällige Ergebnisse zu minimieren?
Wenn es um Fragen der Krankheitsentstehung und (kausale) Einflüsse von Umweltfaktoren geht, ist eine randomisierte Studie häufig schwierig oder sogar unmöglich. So ist z. B. der Effekt von Gewalt oder bekannter Noxen auf die psychische oder körperliche Krankheitsentstehung mittels randomisierter Studien aus ethischen Gründen undenkbar. Nicht-randomisierte (Beobachtungs-) Studien sind daher unabdingbar, da Antworten auf obige Fragen gerechtfertigt sind und auch von klinischer Relevanz. Bei der Interpretation von Beobachtungsstudien ist es wichtig, Fehlerquellen zu kennen, ansonsten besteht die Gefahr, dass wir Zusammenhänge erkennen, welche nur scheinbar kausal sind (Scheineffekte).

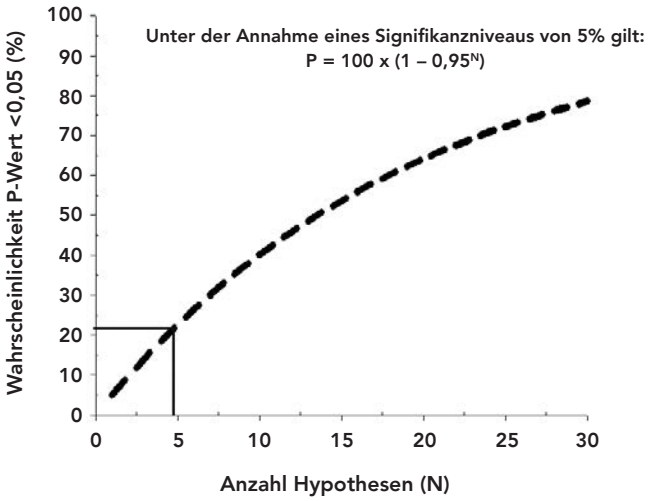

Abbildung 1
Die Abbildung zeigt den Zusammenhang zwischen der Anzahl Subgruppen (N) und der Wahrscheinlichkeit für ein signifikantes Resultat (P-Wert <0,05). Bei 5 Subgruppen ist die Wahrscheinlichkeit 22%, dass in mindestens einer Subgruppe ein signifikantes Resultat gezeigt werden kann.

Kohortenstudie
Die Framinghamstudie ist ein Beispiel epidemiologischer Forschung, welche einen wesentlichen Beitrag zum Verständnis kardiovaskulärer Erkrankungsrisiken geleistet hat [7]. Es handelt sich dabei um eine repräsentative bevölkerungsbasierte Studienpopulation, welche zum Zeitpunkt der Erfassung (damals vermuteter kardiovaskulärer Risikofaktoren «Expositionen») gesund war im Sinne von «keine kardiovaskuläre Erkrankung» vorhanden. Neu aufgetretene kardiovaskuläre Erkrankungen (Inzidenz) wurden zwischenzeitlich über einen Zeitraum von mehr als 30 Jahren detailliert erfasst. Erkrankungs-Inzidenzen in exponierten und nicht-exponierten Bevölkerungsgruppen konnten somit verglichen werden was Rückschlüsse auf kausale Risikofaktoren erlaubte.
Confounding
Wenn wir bei Beobachtungsstudien Zusammenhänge zwischen einer Exposition (X-Variable) und einer Erkrankung (Y-Variable) beobachten, müssen wir uns immer die Frage stellen, ob Störvariablen (Confounder) eine Schein-Kausalität zwischen X und Y vortäuschen. So ist z. B. der signifikante Zusammenhang zwischen «gelben Fingern» (Exposition X) und Lungenkrebs (Y-Variable) das Ergebnis des Confoundings durch Zigarettenkonsum (d. h. der Zusammenhang zwischen «gelben Fingern» und Lungenkrebs kommt dadurch zustande, dass Rauchen ein Risikofaktor für Krebs darstellt und Zigarettenkonsum mit «gelben Fingern» assoziiert ist). *Der kritische Leser von Beobachtungsstudien sollte darauf achten, ob für ein mögliches Confounding kontrolliert wurde.* Häufige Ansätze im Umgang mit Confounding sind:

– Gleiche Verteilung von Fällen und Kontrollen in Bezug auf den vermuteten Confounder (Matching)
– Stratifizierung (d. h. man schaut, ob ein vermuteter Effekt in unterschiedlichen Ausprägungen eines vermuteten Confounders vorkommt, z. B. sollte die Lungenkrebs-Inzidenz in unterschiedlichen Alterskategorien bei den Rauchern immer höher sein als bei den Nichtrauchern, als Hinweis, dass Rauchen unabhängig vom Alter ein Risikofaktor für Krebs darstellt)
– Adjustierung (statistische Modellierung mit Hilfe der Regressionsanalyse)

Die obigen Kontrollmöglichkeiten setzten voraus, dass mögliche Confounder bekannt sind und auch erfasst (gemessen) wurden. *Bei der Interpretation von Beobachtungsstudien kann somit ein «residuelles Confounding» nie ganz sicher ausgeschlossen werden. Der «Goldstandard» für die Kontrolle von Confounding ist die Randomisierung, sie kann garantieren, dass zum Zeitpunkt der Studie unbekannte Confounder zufällig und gleichmässig auf die Gruppen verteilt sind* und damit verzerrte Schlussfolgerungen verhindert werden sollten.

Eine kürzlich im New England Journal of Medicine publizierte Kohortenstudie über die Assoziation zwischen Kaffeekonsum und Mortalität (>300 000 Teilnehmer) illustriert schön, wie sich ein Zusammenhang unter Berücksichtigung möglicher Confounder ändern kann [8]. Vermehrter Kaffeekonsum (>6 Tassen/Tag) war kontrolliert für Alter als wichtigste Confounding Variable für Mortalität ein signifikanter Risikofaktor (Hazard Ratio 1.60; 95%-Vertrauensintervall: 1.51–1.69). Nach zusätzlicher Berücksichtigung von 19 potentiellen Confoundern resultierte in der multivariablen Analyse eine signifikante protektive Assoziation zwischen Kaffeekonsum und Mortalität (Hazard Ratio 0.90; 95%-Vertrauensintervall: 0.85–0.96). Die Autoren schlussfolgern, dass die beobachtete inverse Beziehung zwischen Kaffeekonsum und Mortalität keine Rückschlüsse auf eine kausale Beziehung zulassen.

Fall-Kontroll-Studie

Im Gegensatz zur Kohortenstudie, wo die Erkrankungshäufigkeit zwischen exponierten und nicht-exponierte Personen verglichen wird, werden bei der Fall-Kontroll Studie Expositionshäufigkeiten zwischen erkrankten Personen (Fälle) und Gesunden (Kontrolle) verglichen. Selektions- (Selection Bias) und Informationsfehler (Recall Bias) sind zwei Arten systematischer Fehler, welche den Einfluss einer Exposition auf das Erkrankungsrisiko verfälschen und bei der Fall-Kontroll Studie kritisch beurteilt werden müssen.

Selection Bias

Genauso wie in randomisierten Studien ist es unabdingbar, dass in Fall-Kontroll-Studien die Einschlusskriterien klar definiert sind. Nur so ist eine Beurteilung möglich, ob es sich bei der Auswahl der Fälle um eine repräsentative Population von «Erkrankten» handelt. Zudem sollte man sich im Klaren sein, ob es sich um neu diagnostizierte Fälle (inzidente Fälle) oder prävalente Fälle handelt, da sich bei Inzidenzfällen eher Aussagen über ätiologische Faktoren machen lassen und bei Prävalenzfällen eher Rückschlüsse auf Merkmale des (Langzeit-)Überlebens zulassen. Eine grosse Herausforderung stellt die Auswahl vergleichbarer Kontrollen dar. Der kritische Leser sollte hier insbesondere ein Augenmerk darauf legen, ob hospitalisierte oder nicht-hospitalisierte Personen als Kontrollen rekrutiert wurden.

Recall Bias

Ein Hauptproblem bei Fall-Kontroll Studien ist das Erinnerungsvermögen (Recall). Kritisch ist dabei der Aspekt, dass die Erinnerungsfähigkeit zwischen Kontrollen und Fällen unterschiedlich sein kann. Diese Art der Verzerrung, auch als «Rumination Bias» (Wiederkäuen) bekannt, ist dadurch zu erklären, dass z. B. Mütter, welche ein Kind mit einer Fehlbildung zur Welt gebracht haben, eher versucht sind, ungewöhnliche Ereignisse in der Schwangerschaft mit dem Kind in Verbindung zu bringen, als Mütter von gesunden Kindern. Dieses unterschiedliche Erinnerungsvermögen könnte dann verantwortlich sein, warum im obigen Beispiel ein Zusammenhang zwischen pränatalen (mütterlichen) Infektionen und Fehlbildungen gefunden werden könnte.

Schlussfolgerung

Wenn wir im klinischen Alltag bemüht sind, die beste wissenschaftliche Evidenz in unserer medizinischen Entscheidungsfindung zu berücksichtigen, kommen wir nicht darum herum, wissenschaftliche Arbeiten kritisch auf Validität und Relevanz zu beurteilen. Daher ist es wichtig, dass wir als «Anwender» Kenntnisse über Fehlerquellen und deren Kontrollmöglichkeiten besitzen, welche mit einzelnen Studien Designs verbunden sind. Ein «Screening» auf diese Schwächen kann uns helfen, die «Spreu vom Weizen» zu trennen. Das ursprünglich in Oxford, UK entwickelte Critical Appraisal Skills Programme (CASP) stellt auf ihrer Webside einfache Checklisten für die unterschiedlichen Studien Designs zur Verfügung, welche ein systematisches kritisches Beurteilen wissenschaflticher Evidenz ermöglichen (www.casp-uk.net).

Der Autor hat keine Interessenskonflikte im Zusammenhang mit diesem Beitrag deklariert.

Literatur

1 Sackett DL, Richardson WS, Rosenberg W, et al., editors. Evidence-based medicine: how to practice and teach EBM. London: Churchill-Livingstone; 1997.
2 Hill AB. The Environment and Disease: Association or Causation? Proc R Soc Med. 1965;58:295–300.
3 Gordis L, editor. Epidemiologie: Verlag im Kilian; 2001.
4 Bonita R, Beaglehole R, Kjellström T, editors. Basic epidemiology: World Health Organization; 2006.
5 Boyd CM, Vollenweider D, Puhan MA. Informing evidence-based decision-making for patients with comorbidity: availability of necessary information in clinical trials for chronic diseases. PLoS One. 2012;7:e41601.
6 Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Lancet. 1988;2:349–60.
7 Anderson KM, Wilson PW, Odell PM, et al. An updated coronary risk profile. A statement for health professionals. Circulation. 1991;83:356–62.
8 Freedman ND, Park Y, Abnet CC, et al. Association of coffee drinking with total and cause-specific mortality. N Engl J Med. 2012;366:1891–904.

Korrespondenz:
PD Dr. med. Oliver Senn, MPH
Praxis MediS, Zürich und Institut für Hausarztmedizin
Universität Zürich
Pestalozzistrasse 24
8091 Zürich
oliver.senn[at]usz.ch