Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
URBAN & FISCHER
ZEFQ
www.elsevier.de/zefq
Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) 102 (2008) 253–260

Schwerpunkt

Externe Validität

Jürgen Windeler*

Medizinischer Dienst der Spitzenverbände der Krankenkassen e.V., Essen

Zusammenfassung

Neben der üblichen Qualitätsbewertung von klinischen Studien, die Aspekte der internen Validität betrifft, ist ein anderes Qualitätsmerkmal von Studien, inwieweit ihre Ergebnisse in die Praxis übertragbar sind. Für diesen Aspekt der „externen Validität" gibt es keine ähnlich ausgearbeiteten Prüfinstrumente und Checklisten. Wesentliches Kriterium dieser Qualitätsbewertung ist, ob sich durch die Änderung der Anwendungssituation, die sich entsprechend dem PICO-Schema in verschiedene Aspekte unterteilen lässt, die Effekte einer Therapie ändern. Externe Validität ist also kein Studien-, sondern ein Situationskriterium. Bei der Bewertung geht es nicht darum festzustellen, dass Patienten außerhalb von Studien anders sind als Patienten innerhalb von Studien. Dies ist sicher. Es geht vielmehr darum, ob die Effekte im Sinne eines Unterschieds zwischen zwei Behandlungsgruppen unterschiedlich sind, was man als Effektmodifikation bezeichnet. Generell wird die Beurteilung dadurch erschwert, dass über Effektmodifikationen und deren Einflussfaktoren wenig bekannt ist. Die Bewertung der externen Validität ist daher eher eine Frage fachlichen Ermessens, gestützt auch auf pharmakologische und biologische Informationen.

Schlüsselwörter: Klinische Studien, externe Validität, Übertragbarkeit, Effektmodifikation

External Validity

Summary

It is widely accepted that clinical trials have to be carefully reviewed for internal validity. In addition, aspects of external validity, which is also known as 'generalizability' or 'directness', must be considered. The question of whether the study results can be applied to clinical practice under different conditions than the study itself is of major importance. In contrast to internal validity, external validity has to be viewed as an aspect of the situation, not of a study per se. Assessment of external validity addresses the question of whether effects (comparisons between treatments) are different between patient groups or clinical situations. It is not sufficient and may even not be important whether the patients differ. In epidemiology this situation is well known as 'effect modification.' External validity can be assessed according to the PICO scheme. However, empirical data about effect modifiers are scarce. Consequently, external validity is merely a matter of clinical judgement.

Key words: clinical trials, external validity, generalizability, effect modification

Die Diskussion um die Qualität von Studien und deren Aussagen wird seit Jahrzehnten intensiv geführt. Sie hat zur Entwicklung aussagefähiger Designs beigetragen und es wurden Studien durchgeführt, in denen wichtige Instrumente zum Schutz vor systematisch verfälschten Aussagen (Bias) implementiert wurden. Die Einführung einer evidenz-basierten Medizin hat dazu beigetragen, diese Diskussion zu beleben und zu strukturieren. Sie hat zur

*Korrespondenzadresse: Prof. Dr. med. Jürgen Windeler, Medizinischer Dienst der Spitzenverbände der Krankenkassen e.V., Lützowstraße 53, 45141 Essen.
Tel.: 0201 8327125; fax: 0201 8327402.
E-Mail: j.windeler@mds-ev.de

Verbreitung der Hierarchie von Studiendesigns (Evidenzklassen) geführt und zu zahlreichen Checklisten, die die Qualität von Studien strukturiert abbilden sollen [1].

Die Qualität, die mit diesen Bewertungsinstrumenten erfasst werden soll, stellt aber nur einen, wenn auch wesentlichen Teil der notwendigen Qualitätsbewertung von Studien dar. Fehler können aber nicht nur bezüglich der Fragen auftreten, die sich die Studien zur Beantwortung gestellt haben. Sie sind auch möglich bei der Übertragung von Studienergebnissen in die Praxis. Diese beiden Dimensionen der Qualitätsbewertung bezeichnet man als interne bzw. als externe Validität. Für letztere sind auch die nicht ganz eindeutigen und auch nicht ganz identischen Begriffe „Verallgemeinerbarkeit", „Übertragbarkeit", "generalizability" [2,3] oder "directness" [4] in Gebrauch.

Die anfangs beschriebene Qualitätsbewertung betrifft ausschließlich Aspekte der internen Validität. Diese sind umfassend, wenn auch sicher nicht abschließend, erforscht. Publikationen zu dieser Thematik sind Legion. Wesentliche Eckpunkte sind in offiziellen Dokumenten wie z. B. Guidelines der internationalen Regulierungsbehörden niedergelegt. Qualitätsaspekte sind in Checklisten abgebildet und ihre Bewertung teils standardisiert. Schließlich kennt jeder die vielen Beispiele, in denen Studien von unterschiedlicher interner Validität zu unterschiedlichen Ergebnissen geführt haben und in denen man durch intern valide Studien systematische Fehler, die aufgrund der Ergebnisse anderer Studien begangen wurden, korrigieren konnte [5]. Aktuellste Beispiele sind zweifellos die Geschichte der Hormontherapie und die der Vitaminsubstitution [6–8].

Im Unterschied zu diesem sehr intensiv bearbeiteten Thema herrscht bei der Frage der externen Validität weitgehende Leere. Es gibt keine systematische Bearbeitung, keine Checklisten und keine Standards zur Bewertung. Es gibt, wenn man es genau betrachtet, nur wenige gute, empirisch belegte Beispiele dafür, inwieweit hier Fehler auftreten können. Und es gibt, teilweise möglicherweise als Konsequenz dieser Leere, vielfältige Informationsdefizite und überwiegend auf Eindrücken und Kolportage beruhende Diskussionen.

Man könnte zu dem Schluss kommen, dass die externe Validität von geringerer Bedeutung ist als die interne Validität. Eigentlich ist aber das genaue Gegenteil der Fall. Vergegenwärtigen wir uns die Situation einer klinischen Studie (Abb. 1):

Die Frage, die eigentlich beantwortet werden soll, betrifft eine Zielpopulation, d.h. Patienten, die zu irgendeinem zukünftigen Zeitpunkt behandelt werden sollen. Um diese Frage zu beantworten, definiert man eine Studienpopulation, die nicht mit der Zielpopulation identisch ist und sein kann. Abgesehen von speziellen Notwendigkeiten einer Studie (Risikominimierung durch Ausschluss von Kontraindikationen, besondere Diagnoseabsicherung, informed consent) ist banaler Weise zu berücksichtigen, dass Studienpatienten im Jahr 2008 rekrutiert werden, die Ergebnisse dieser Studie aber für Patienten im Jahre 2013 relevant sind. Man wird jedenfalls feststellen müssen, dass die Studienpopulation weder mit der Zielpopulation identisch ist noch ein Ausschnitt dieser Zielpopulation darstellt. Entsprechend den Definitionen der Studienpopulation werden die Studienpatienten für eine Studie ausgewählt und für die Studie rekrutiert, behandelt und beobachtet. Nehmen wir an, dass Ergebnis sei valide im Sinne einer hohen internen Validität. Das entscheidende Problem ergibt sich dann.

Eine *statistische* Schlussfolgerung ist nur von den Studienpatienten auf die Studienpopulation möglich, nicht jedoch auf die Zielpopulation. Der mögliche Schluss ist jedoch eigentlich irrelevant. Zum einen steht ja die Zielpopulation im Fokus des Interesses im Sinne einer Entscheidung über die Behandlung zukünftiger Patienten. Da aber die Studienpopulation nicht der Zielpopulation entspricht, ist ein Schluss von der Studienpopulation auf die Zielpopulation nicht ohne weiteres möglich. Zum anderen existiert die Studienpopulation in der Realität gar nicht, sondern nur über die sie definierenden Kriterien. Ein Schluss von den Studienpatienten auf die Studienpopulation hat also, streng genommen, wenig Bedeutung für real existierende Patienten, die das zentrale Interesse einer jeden Studie sind. Der eigentlich relevante Schluss ist der von den Studienpatienten (bzw. der Studienpopulation) auf die Zielpopulation, dieser ist jedoch nicht durch statistische Methoden zu unterstützen, sondern er ist eine medizinische Bewertung, eine Frage des inhaltlichen Ermessens. Damit reichen für dieses Urteil nicht die Studienergebnisse selbst aus, sondern es müssen, soweit verfügbar, ergänzende Informationen, empirische Daten und/oder biologische, physiologische und pharmakologische Erkenntnisse und Modelle herangezogen werden.

Betrachtet man den eigentlichen Zweck von klinischen Studien, so ist nachvollziehbar, dass die Schlussfolgerung von den Studienpatienten auf die Zielpopulation entscheidend und für die Umsetzung von Studienergebnissen einzig relevant ist. Hervorzuheben ist jedoch, dass es wenig sinnvoll ist, ohne sichergestellte interne Validität einer Studie über die externe Validität nachzudenken und diese zu beurteilen. Es

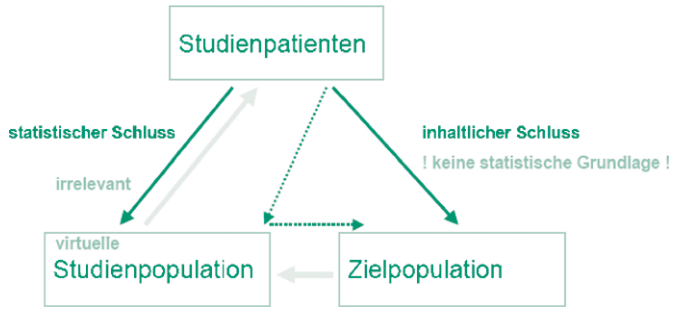

Abb. 1. Beziehung zwischen Zielpopulation und Studienpatienten

mag zwar so sein, dass sog. „Anwendungsbeobachtungen" von Arzneimitteln, unmittelbar in der praktischen Versorgung durchgeführt, einen sehr guten und praktisch relevanten Einblick in die Versorgung liefern. Da sie aber mangels interner Validität etwa für die Bewertung des Nutzens von Arzneimitteln ungeeignet sind, ist die Frage der externen Validität zwar zu bejahen, mangels Voraussetzung aber bedeutungslos.

Effektmodifikation

Regelhaft wird über die Frage der externen Validität mit Argumenten diskutiert, die nicht oder jedenfalls nicht vollständig zutreffend sind. Ein entscheidendes Problem sei an dem folgenden Beispiel erläutert. In einer sehr verdienstvollen Studie haben Brown et al. [9] die Prognose von Patienten mit Herzinfarkt unter verschiedenen Therapien untersucht. Abbildung 2 zeigt die Überlebenskurven von 3 Patientengruppen, oben die Patienten, die eine thrombolytische Therapie im Rahmen einer Studie erhalten haben, darunter die Patienten, die eine thrombolytische Therapie außerhalb der Studie erhalten haben. Ganz unten liegt die Kurve für diejenigen Patienten, die keine thrombolytische Therapie erhalten haben, typischerweise deshalb, weil sie für eine solche Therapie als nicht geeignet angesehen wurden. Das Problem liegt nun nicht in dieser Studie und ihrer Darstellung, sondern in der Interpretation der Autoren. Sie schreiben: "Our findings raise doubts about the validity of the evidence base for thrombolytic treatment for the generality of patients with myocardial infarction".

Diese Interpretation ist aber nicht zutreffend. Es geht bei der Frage der Übertragbarkeit nicht darum, ob Patienten in Studien „anders" sind als Patienten in der späteren Praxis (das ist sicher), sondern es geht darum, ob die Therapieeffekte anders sind. Diese sind definitionsgemäß Unterschiede zwischen der Anwendung und der Nichtanwendung einer Intervention, also nicht Vorher-Nachher-Vergleiche. Im konkreten Beispiel heißt dies, dass die "evidence base" so lange unproblematisch ist, wie nicht die „Effekte" in den 3 Gruppen unterschiedlich sind (wobei dies für die unterste Gruppe noch nicht einmal gilt, denn es erschließt sich nicht ohne weiteres, warum die evidence base für eine thrombolytische Therapie für Patienten relevant sein sollte, die als für diese Therapie nicht geeignet angesehen werden). Ein Vergleich der oberen beiden Kurven sagt keineswegs, dass Patienten in klinischen Studien größere Therapieeffekte haben als Patienten außerhalb der Studien. Der Vergleich wird bezüglich der „Relevanz der Ergebnisse von klinischen Studien für die medizinische Praxis" erst aussagekräftig, wenn die jeweiligen Vergleiche (Behandlung dieser Patienten ohne Thrombolyse) ergänzt und bewertet werden könnten.

In Tabelle 1 wird das Problem der sogenannten „Effekt-Modifikation" illustriert: Stellt man sich zwei Behandlungssituationen mit Basisrisiken von 10% und 1% für ein ungünstiges Ergebnis vor, so könnte eine Therapie dieses Risiko in beiden Situationen halbieren. Das relative Risiko (RR) betrüge dann in beiden Situationen 0,5 (Spalten A und B). Lässt man ergänzend notwendige Betrachtungen über Schäden vereinfachend außer Acht, so würde man feststellen, dass die Therapieeffekte in beiden Therapiesituationen gleich sind. Die Aussage zum Nutzen einer Behandlung wäre daher übertragbar. Zu erörtern ist die Frage, ob für diese Bewertung der Effektmodifikation das relative Maß oder das absolute Maß betrachtet werden soll. Dies ist relevant, denn es ist, wie aus der Tabelle ersichtlich, für das absolute Maß eine deutliche Modifikation des Effektes (Faktor 10) festzustellen. Es gibt Untersuchungen, die darauf hinweisen, dass relative Therapieeffekte bei verschiedenen Basisrisiken konstant sind [10]. Die Folge (s.o.) ist zwar, dass die absoluten Effekte nicht konstant sein können, wegen des konstanten relativen Effektes ergäbe sich aber die Situation, dass die absoluten Ergebnisse auf andere Basisrisiken einfach umgerechnet werden könnten, was eine Übertragung ermöglichen würde.

Die Spalten C und D der Tabelle zeigen eine (möglicherweise sehr artifizielle) Situation, in der umgekehrt die Effekte im absoluten Maß gleich und im

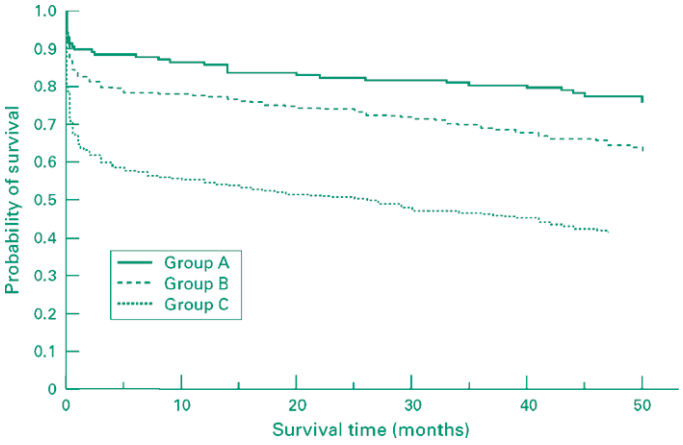

Abb. 2. Überlebenskurven dreier Patientengruppen nach Myokardinfarkt. A: Thrombolyse innerhalb von Studien; B: Thrombolyse außerhalb von Studien; C: keine Thrombolyse (nach Brown et al. [9])

Tabelle 1. Illustration einer Effektmodifikation; ACHTUNG: die vier Zahlenspalten müssen bitte mit A, B, C und D betitelt werden

Ereignisrate Kontrolle (%)	10	1	10	1
Ereignisrate Intervention (%)	5	0,5	9,5	0,5
RR	0,5	0,5	0,95	0,5
ARR (%-Punkte)	5	0,5	0,5	0,5
NNT	20	200	200	200

relativen Maß unterschiedlich sind. Wegen der starken Abhängigkeit der absoluten Risikoreduktion von dem Basisrisiko könnte es sinnvoll sein, Effektmodifikation primär auf der Basis der relativen Maße oder der standardisierten absoluten Risikoreduktion (Differenz geteilt durch Standardabweichung) zu beurteilen.

Ergänzend zu dem wesentlichen Hinweis, dass es sich bei der Frage der externen Validität nicht um die Frage handelt, ob die Patientencharakteristika (inkl. deren Basisrisiko) unterschiedlich sind, sondern dass es um einen Unterschied in den Effekten geht, sind noch zwei weitere Bemerkungen von Bedeutung: Bei der externen Validität geht es weniger als bei der internen Validität um die Vermeidung von „Bias" sondern um die Abschätzung, ob der Fehler sehr groß oder eher hinnehmbar ist (streng genommen ist der Begriff Bias hier unangebracht, da, wie oben ausgeführt, die Studienpatienten eben keine Stichprobe der Zielpopulation sind und damit eine statistische Schlussfolgerung auf die Zielpopulation gar nicht möglich ist).

Zum Zweiten sei darauf aufmerksam gemacht, dass der Begriff „externe Validität" nicht mit anderen Begriffen, z. B. der Übertragbarkeit von Ergebnissen gleichgesetzt werden sollte. Externe Validität wird, wie der Begriff sagt, typischerweise auf Patienten bezogen, die nicht die Charakteristika der Studienpopulation erfüllen. Es kann aber auch bei der Übertragbarkeit von Studienergebnissen auf solche Patienten ein Problem geben, die Teil der Studienpopulation (und damit der Studienpatienten) sind, etwa bei der Frage der Behandlung von Untergruppen der Studienpopulation. Diese Frage ist aber nicht die Frage nach der externen Validität im engeren Sinne, sondern die nach Aussagen über Subgruppen, für die eine gesonderte methodische Diskussion erforderlich ist [11,12].

Bewertung der externen Validität

Ein häufig verfolgter Weg zur Bewertung der externen Validität ist, den Anteil der Patienten als Kriterium heranzuziehen, die von allen Patienten mit einer Indikation in klinischen Studien abgebildet werden. Für Studien zur Therapie von Carotis-Stenosen und für Schlaganfallpatienten ist z. B. beklagt worden, dass weniger als 1% dieser Populationen in den Studien vertreten sind. Einstellige Prozentsätze sind vermutlich für die meisten Indikationen anzutreffen. Bei dieser Argumentation ist aber daran zu erinnern, dass es nicht darauf ankommt, dass bestimmte Patienten in den Studien vertreten sind, sondern darauf, ob die Therapieeffekte bei diesen Patienten anders sind (oder ob es gute Gründe gibt, dass sie anders sein könnten) als bei den Patienten, die in der Studie vertreten sind. Entsprechende Daten oder auch belastbare Argumente liegen oft nicht vor, so dass die Relevanz der genannten %-Werte schwer abzuschätzen ist.

Eine ganz andere Frage ist, ob bestimmte Patientengruppen, von denen man weiß oder aufgrund guter Argumente annehmen kann, dass die Therapieeffekte anders sind, in Studien überhaupt vertreten sind. Insbesondere bzgl. der Risikobewertung kann es in solchen Fällen geboten sein, die externe Validität nicht als gegeben anzunehmen. Als Faustregel wird man davon ausgehen müssen, dass das Risiko der Anwendung „im Alltag" gegenüber Studienbedingungen erhöht ist (s. das Beispiel zur Behandlung mit Spironolakton in [2]).

Ein anderer Weg, die externe Validität zu beurteilen, könnte in einem Vergleich von Ergebnissen randomisierter und nicht-randomisierter Studien liegen, nämlich dann, wenn man der These folgt, dass nicht-randomisierte Studien generell eine hohe oder ausreichende externe Validität aufweisen. Zu diesem Vergleich gibt es eine ganze Reihe von Übersichten und Metaanalysen, die in einem Cochrane-Review 2007 [13] zusammengefasst worden sind. Das Ergebnis dieser Vergleiche hilft aber in der Bewertung der externen Validität nicht wirklich weiter. Zum Einen ist die Frage der Referenz unklar, also, ob die randomisierten oder nicht-randomisierten Studien jeweils als extern valide anzusehen sind. Zum Anderen sind diese Vergleiche selbst sicher nicht intern valide (und bringen noch andere methodische Probleme mit sich, z. B. ein Power-Problem). Eine neuere Studie [14] zeigt zudem, dass – bei gleicher Indikation und Rahmenbedingungen – randomisierte und nicht-randomisierte prospektive vergleichende Studien dann gleiche Ergebnisse liefern, wenn beide (!) eine sehr gute Qualität, also eine hohe interne Validität aufweisen. Das Defizit der Vergleiche würde durch den Vorschlag sogenannter „meta-randomisierter Vergleiche" behoben werden können, für die es aber bisher keine praktischen Beispiele gibt [15].

Für eine Bewertung der externen Validität kursiert des weiteren ein Vorschlag, der auf der These basiert, dass die externe Validität komplementär zur zunehmenden internen Validität in typischen Evidenzhierarchien abnimmt. Diese Idee ist aber offensichtlich falsch. Die interne Validität ist das Sortierungskriterium für solche Evidenzhierarchien. Da die externe Validität eine ganz andere Bewertungsdimension betrifft, ist nicht einsichtig, warum sich „zufällig" ein komplementäres Bild ergeben sollte. Auch im Konkreten ist die Auffassung, dass Fallserien generell eine höhere externe Validität aufweisen als z. B. RCT, nicht haltbar, da ohne Zweifel eine Fallserie von Patienten aus Zentralafrika für die Mecklenburgische Versorgung weniger Relevanz haben wird als ein RCT in Mecklenburg. Wenn überhaupt, so wäre die externe Validität jeweils innerhalb einer Evidenzstufe zu bewerten, aber auch diese Systematik verkennt einen ganz entscheidenden Punkt: Während die interne Validität ein Studiencharakteristikum ist (man kann sagen: eine Studie ist intern valide), ist die externe Validität ein Situationskriterium (man kann sagen: diese Studie ist für meine Anwendungssituation extern valide, sie muss damit aber nicht für eine andere Anwendungssituation extern valide sein). Es ist aus diesem Grund auch wenig zielführend, Checklisten für externe Validität [16] zu entwickeln, die Studien in diesem Sinne eine bestimmte hohe Qualität bescheinigen sollen.

Betrachtet man diese Aussage am Beispiel einer Studie, die bei 60–80-jährigen

Männern in der stationären Versorgung durchgeführt wurde, so wird man (Tabelle 2) die Übertragung auf 55-jährige Männer in aller Regel unproblematisch bewerten. Die Übertragung auf 60–80-jährige Frauen wird von der spezifischen Problemstellung und der Indikation abhängen, also bei der Behandlung von Harnwegsinfekten anders zu bewerten sein als bei grippalen Infekten. Die Frage der Übertragung von Ergebnissen aus der stationären Versorgung wird man für andere Krankenhäuser relativ unproblematisch, auch für viele Situationen in die ambulante Versorgung für unkompliziert, dagegen von der Tertiär- in die Primärversorgung oft als schwierig bis unmöglich betrachten (zur Erinnerung: relevant ist dabei immer die Frage, ob *Therapieeffekte* unterschiedlich sind). Dagegen wird man sich bei der Frage der Übertragung von Studiendaten von älteren Erwachsenen für Kleinkinder grundsätzlich schwer tun.

Bewertungskriterien PICO et al.

Betrachtet man etwas genauer, von welchen Einflussfaktoren die Frage der externen Validität abhängt, als wie bedeutsam diese Faktoren zu bewerten sind und in inwieweit sich an diesen Faktoren etwas ändern ließe, inwieweit also die externe Validität zu optimieren ist, so bietet sich eine Erweiterung des bekannten PICO-Schemas an (patient, intervention, control, outcome). Auch diese Betrachtung hat zur Grundlage, dass nicht die Frage betrachtet wird, ob sich Studienpatienten und Zielpopulationen in bestimmen Charakteristika mehr oder weniger gravierend unterscheiden, sondern die einzig relevante Frage, ob diese Charakteristika als Effektmodifikatoren, also als Faktoren, die die Größe der Effekte, also das Ausmaß des Nutzens beeinflussend gelten können und zu berücksichtigen sind.

- Patienten valider Studien und die Zielpopulation können sich in den Patienteneigenschaften selbst unterscheiden. Zu nennen ist das Alter, das Geschlecht, das Krankheitsstadium oder auch Risikokonstellationen, Multimorbidität und Kontraindikationen. So selbstverständlich es einerseits ist, dass unter den Bedingungen von Studien Indikationen zuverlässig gestellt und Kontraindikationen beachtet sein müssen, so wenig hilfreich ist es meistens, sehr strenge Einschluss- und Ausschlusskriterien zu definieren. Mindestanforderung ist es aber, diese Einschlusskriterien genau in einer Studienpublikation zu beschreiben, um dem Leser die Möglichkeit zu geben, die externe Validität der Studie für seine Anwendungssituation zu bewerten.
  Die Relevanz der Patientenmerkmale ist schwierig zu beurteilen. Es existieren abgesehen von einigen guten Beispielen nur wenige Daten dazu, welche Effektmodifikation durch diese Charakteristika hervorgerufen wird. Hier ist am ehesten das Krankheitsstadium zu berücksichtigen, möglicherweise bestimmte genetische Unterschiede und in besonderen Einzelfällen das Geschlecht.
- In Erweiterung des PICO-Schemas sind zum zweiten Charakteristika zu nennen, die nur wegen der Durchführung einer Studie auftreten. Dieses ist z. B. das Einverständnis der Patienten (in die Studie!), die dafür notwendige spezifische Aufklärung, eine möglicherweise besondere Erwartungshaltung, die durch die besonderen Bedingungen einer Studie (positiv?) bzw. die Aussage des betreuenden Arztes zur Zufallsauswahl (negativ?), ggf. auch durch die Verblindung beeinflusst werden kann.
  Die Relevanz dieser studienspezifischen Merkmale ist sehr schwer zu beurteilen. Es gibt einige wenige Untersuchungen [17] aus Zeiten, in denen eine aufgeklärte Einwilligung aller Patienten für die Studien noch nicht als zwingend angesehen worden ist, die Einflüsse der Studienbedingungen auf Therapieeffekte beschreiben. Geklärt ist hier wenig, zu ändern sind diese Bedingungen in üblichen Studiendesigns nicht.
- Als zweite Ergänzung des PICO-Schemas sollten spezifische Prozeduren erwähnt werden, die nur innerhalb der Studie bzw. für die Studie durchgeführt werden. Besonders zu nennen sind hier sogenannte Run-in-Perioden, in denen z. B. alle Patienten zunächst mit Verum behandelt und nur die "Responder" danach randomisiert werden. Umgekehrt gibt es ein Vorgehen, in dem alle Patienten zunächst mit Placebo behandelt werden und alle "Non-Responder" randomisiert werden. Diese Designs dienen einer künstlichen Homogenisierung der Studienpatienten. Das Vorgehen hat aber mit dem Alltag nichts zu tun, wo diese Run-in-Perioden (natürlich) nicht erfolgen. Ähnliches gilt für Vorgehensweisen, in denen Vergleichstherapien unangemessen durchgeführt oder falsch dosiert werden.
  Selbstverständlich kann in einer Studie anders verfahren werden. Man wird die beschriebenen Vorgehensweisen als K.O.-Kriterien in dem Sinne bezeichnen müssen, als diese Studien für keine praktisch vorstellbare Situation extern valide sind. Da das Ziel solchen Vorgehens gerade eine Effektmodifikation ist, also, Therapieeffekte künstlich zu vergrößern, spricht dies bereits dafür, dass hier eine relevante Effektmodifikation vorliegt und die externe Validität nicht gegeben ist. Für eine solche Situation wäre in Umkehrung des üblichen Vorgehens der Nachweis zu fordern, dass die Therapieeffekte übertragbar sind.
- Als Interventionseigenschaften sind etwa die Therapiedosierung, Begleittherapien, die Applikationsart

Tabelle 2. externe Validität eines Studienergebnisses bei 60-80jährigen Männern in der stationären Versorgung für andere Anwendungssituationen

Manner 55 J	Frauen 60–80	Praxis	Kleinkinder
++	++bis−−	+bis−	−−

oder auch das Training der Behandler zu nennen. Die Relevanz wird im Einzelfall zu prüfen und wohl nicht selten zu bejahen sein. Man wird die Bewertung in der Regel auf der Basis pharmakologischer, physiologischer und biologischer Kenntnisse und Modelle durchführen müssen. Dass solche Bewertungen durchaus problematisch sein können, zeigt ein Blick auf die (relative) Niedrig-Dosis-Behandlung mit Acetylsalicylsäure, wo im Bereich einer Vervierfachung der Dosis (70–300 mg) kein Unterschied in der Wirksamkeit (wohl aber des Blutungsrisikos) beobachtet werden konnte [18].
Grundsätzlich lässt sich diese Problem der externen Validität dadurch angehen, dass Kernpunkte der späteren praktischen Durchführung in der Studie berücksichtigt werden.
- Als Eigenschaften der Outcome-Messung sind z. B. Surrogatkriterien, validierte Skalen, klinische Endpunkte zu nennen, aber auch die Zeit der Beobachtung. Auch hier ist zur Frage der externen Validität bezüglich der *gleichen Endpunkte* (sind die Effekte in den Surrogatkriterien in der Zielpopulation anders?) wenig bekannt. Die Frage, ob Effekte aus Surrogatkriterien sich tatsächlich in klinisch relevanten Endpunkten abbilden, ist genau genommen keine Frage der externen Validität.

Wesentlich dafür, die externe Validität überhaupt bewerten zu können ist es aber, dass die Details der Studie entlang des PICO-Schemas vollständig dargestellt werden [19].
Speziell bei der Betrachtung der Interventionseigenschaften (s.o.) wird eine weitere allgemeine Frage offenkundig: Bis zu welcher Abweichung der Behandlung von der Studientherapie kann man noch von dem Problem externer Validität sprechen? Anders ausgedrückt: Handelt es sich bei wesentlichen Änderungen in diesen Eigenschaften eigentlich noch um den Unterschied zwischen Studie und Alltag oder nicht eigentlich um eine neue, eigene Therapiemethode? Halbiert man etwa im „Alltag" gegenüber der Studie die Dosis eines Medikamentes, so wird man dies nicht ohne weiteres unter dem Stichwort „externe Valdität" betrachten können, ebenso wenig, wenn man Studienergebnisse aus hochkompetenten Zentren auf Praxisanfänger überträgt. Die Grenzziehung ist sehr schwierig (siehe Aspirinbeispiel), zumal die hierfür erforderlichen Kenntnisse (welche Auswirkungen haben bestimmte Abweichungen) meistens fehlen.

Fazit

Die Ergebnisse von Studien sind vorrangig nicht für die Studienpatienten, sondern für die Entscheidungen bei anderen, zukünftigen Patienten von Bedeutung. Damit hat auf der Basis intern valider Studienergebnisse der Aspekt der externen Validität für die klinische Forschung zentrale Bedeutung. Dieser eindeutigen „qualitativen" Bedeutung steht allerdings eine nicht so klare „quantitative" Bedeutung gegenüber. Diese erschließt sich grundsätzlich nicht über die Frage, ob Patienten in klinischen Studien „anders" sind als diejenigen, die von den Studienergebnissen betroffen sind, sondern sie wird durch die Frage abgebildet, ob die Effekte bei anderen Patienten relevant anders sind als bei den Studienpatienten. Man muss leider feststellen, dass es zu dieser Frage wenige methodisch belastbare Daten und Aussagen gibt. Oft betreffen solche Ergebnisse erhöhte Risiken, die durch Änderung der Patientenpopulation (Stichwort z. B. Multimorbidität) oder durch Änderung der Indikationsstellung oder der Therapiekontrolle auftreten können.

Insgesamt lassen sich zur externen Validität zur Zeit folgende Aussagen machen:

- Die Übertragung von Studienergebnissen auf Patienten/Situationen kann zu systematischen Fehlern führen. Solche Fehler lassen sich leicht exemplarisch benennen, wie häufig und wie relevant sie sind, ist aber nicht bekannt.
- Externe Valdität steht für eine mögliche Effektmodifikation in anderen Situationen. Sie ist damit ein Situations-, kein Studienkriterium und lässt sich durch studienbewertende Ansätze nicht abbilden.
- Studien sollten keine Maßnahmen implementieren, die in der Praxis nicht eingesetzt werden. Als ein Beispiel sind sogenannte Run-in-Perioden zu nennen.
- Es ist sinnvoll, Berichtsstandards zu Studiendetails zu fordern, die für die Bewertung der externen Validität wichtig sein können.
- Primär wird es um eine fachliche, situationsbezogene Bewertung gehen, die nach der Vorstellung evidenzbasierter Medizin in der Regel dem Ausschlussprinzip folgt.

Wesentliche Aspekte externer Validität sind unbekannt und Aussagen daher spekulativ. Die Forderung nach mehr Forschung, die vor allem die Identifizierung relevanter Effektmodifikatoren zum Ziel haben müsste, ist richtig und ein sehr viel versprechendes Feld für die Versorgungsforschung. Die Herausforderung von Studienplanungen liegt darin, einzelne mögliche Effektmodifikatoren zu analysieren und damit Beiträge zur Aufklärung wesentlicher Einflussfaktoren für die externe Validität zu liefern. Der Ruf nach „Alltag" hilft nicht.

Literatur

[1] Moher D, et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62–73.
[2] Weiss NS, Koepsell TD, Psaty BM. Generalizability of the results of randomized trials. Arch Intern Med 2008;168:133–5.
[3] Bailey KR. Generalizing the results of randomized clinical trials. Contr Clin Trials 1994;15:15–23.
[4] GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004Jun19;328(7454):1490.
[5] Abel U, Windeler J. Irrtümer in der Bewertung medizinischer Therapien. Ursachen und Konsequenzen. Internist Praxis 1995; 35:613–29.
[6] Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women:

principal results from the Women's Health Initiative randomized controlled trial. JAMA 2002;288:321–3.
[7] Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis. JAMA 2007; 297:842–57.
[8] The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N Engl J Med 1994 Apr 14;330(15):1029–35.
[9] Brown N, Melville M, Gray D, Young T, Skene AM, Wilcox RG, et al. Relevance of clinical trial results in myocardial infarction to medical practice: comparison of four year outcome in participants of a thrombolytic trial, patients receiving routine thrombolysis, and those deemed ineligible for thrombolysis. Heart 1999Jun;81(6): 598–602.
[10] Boissel JP, Cucherat M, Nony P, Chabaud S, Gueyffier F, Wright JM, et al. New insights on the relation between untreated and treated outcomes for a given therapy effect model is not necessarily linear. J Clin Epidemiol 2008;61:301–7.
[11] Rothwell PM. Treating individuals: Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86.
[12] Janatzek S. Aussagekraft von Subgruppen-Analysen, 2004. 〈http://www.mds-ev.org/download/subgruppen-gutachten.pdf〉.
[13] Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. Cochrane Library 3 (2007).
[14] Furlan AD, Tomlinson G, Jadad AR, Bombardier C. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. J Clin Epidemiol 2008;61: 209–31.
[15] Kuss O, Legler T, Börgermann J. Gibt es einen Unterschied zwischen randomisierten und nicht-randomisierten Studien? Evidenz aus einer "Meta-Propensity Score-Analyse" in der Herzchirurgie 54. Biometrisches Kolloquium, München, 10.–13.3.2008.
[16] Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040–8.
[17] Kaptchuk TJ. The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? J Clin Epidemiol 2001; 54:541–9.
[18] Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. BMJ 2002;324: 71–86.
[19] Rothwell PM. Treating Individuals: External validity of randomised controlled trials: "To whom do the results of this trial apply"? Lancet 2005;365:82–93.

Magazin

WHO-Projekt "Action on Patient Safety: High 5s" – Hintergrundinformationen zur Teilnahme Deutschlands

Hintergrund

Das Aktionsbündnis Patientensicherheit (APS) und das Ärztliche Zentrum für Qualität in der Medizin (ÄZQ) sind für die Durchführung eines Vorprojekts zu dem internationalen WHO-Projekt „High 5s" in Deutschland verantwortlich. Daher sucht das ÄZQ Krankenhäuser, die sich für die Teilnahme an diesem zukunftsweisenden Projekt interessieren.

Das internationale WHO-Projekt "Action on Patient Safety: High 5s"

Ziel des Projekts der World Alliance for Patient Safety (WHO) ist die Verbesserung der Patientensicherheit auf internationaler Ebene. Diesbezüglich wurden fünf Themenbereiche – die "High 5s"-als besonders dringlich bewertet. Im Rahmen des Projekts werden zu jedem dieser Themen standardisierte Handlungsempfehlungen ("SOPs" = Standard Operating Protocols) erarbeitet:

- Management von konzentrierten injizierbaren Medikamenten (Managing Concentrated Injectable Medicines)
- Sicherstellung der richtigen Medikation bei Übergaben im Behandlungsprozess (Assuring Medication Accuracy at Transitions in Care)
- Kommunikation bei Übergaben im Behandlungsprozess (Communication During Patient Care Handovers)
- Verbesserte Handhygiene zur Vermeidung krankenhausassoziierter Infektionen (Improved Hand Hygiene to Prevent Health-Care-Associated Infections)
- Vermeidung von Eingriffsverwechslungen (Performance of Correct Procedure at Correct Body Site)

Diese Handlungsempfehlungen sollen zunächst im Rahmen eines Pilottests in 10 Krankenhäusern pro Teilnehmerland eingeführt werden; die Implementierung wird begleitet und evaluiert. Im Anschluss daran ist eine möglichst flächendeckende Einführung geplant. Die Gesamtlaufzeit des Projekts beträgt 4 bis 5 Jahre.
Die Finanzierung des Projekts erfolgt multinational durch die teilnehmenden Staaten und mit Unterstützung des Commonwealth Funds. Das WHO Collaborating Centre for Patient Safety, dessen Aufgaben von der Joint Commission und der Joint Commission International (JCI) wahrgenommen werden, koordiniert das Projekt. Außer Deutschland nehmen die Niederlande, Großbritannien, die USA, Kanada, Australien und Neuseeland an dem Projekt teil.

"High 5s" in Deutschland

In jedem der teilnehmenden Länder ist eine Lead Technical Agency für die Koordinierung des Projekts zuständig. In Deutschland ist dieses das Aktionsbündnis Patientensicherheit (APS), das Ärztliche Zentrum für Qualität in der Medizin (ÄZQ) leistet die operationale Durchführung.

APS und ÄZQ

Das Aktionsbündnis für Patientensicherheit (APS) und das Ärztliche Zentrum für Qualität in der Medizin (ÄZQ) setzen sich seit mehreren Jahren gemeinsam für eine sichere Gesundheitsversorgung in Deutschland ein. Es wurden bereits eine Reihe von Projekten zur Erforschung, Entwicklung und Verbreitung dazu geeigneter Methoden begleitet (z. B. Herausgabe von Empfehlungen für Krankenhäuser, Entwicklung von Trainingsmaterialien, etc.). Weitere Informationen über die Tätigkeiten von APS und ÄZQ finden sich unter www.aktionsbuendnis-patientensicherheit.de, www.forum-patientensicherheit.de und www.azq.de.