Individuelle Therapie 1

Externe Validität randomisierter kontrollierter Studien: „Auf wen sind die Ergebnisse dieser Studie anwendbar?“

Peter M. Rothwell

Stroke Prevention Research Unit, University Department of Clinical Neurology, Radcliffe Infirmary, Oxford OX2 6HE, UK

aus: *Lancet* 2005; **365**: 82–93: (Originaltitel: „To whom do the results of this trial apply?“)

Zusammenfassung

Wenn Ärzte und Patienten Therapieentscheidungen treffen, müssen sie dabei die Ergebnisse relevanter randomisierter kontrollierter Studien (RCTs) und systematischer Übersichtsartikel (Reviews) berücksichtigen. Ihre Relevanz hängt von der externen Validität (oder Generalisierbarkeit) ab, d. h. von der Frage, ob sich die Ergebnisse im Praxisalltag sinnvoll auf eine definierbare Gruppe von Patienten in einem bestimmten klinischen Setting übertragen lassen. Ärzte äußern sich besorgt über die oftmals mangelhafte externe Validität, vor allem von Studien der pharmazeutischen Industrie – eine Beobachtung, die dazu geführt hat, dass wirksame Therapien zu selten eingesetzt werden (sog. „Underuse“). Und doch wird die externe Validität von Forschern, Studienträgern, Ethikkommissionen, der pharmazeutischen Industrie, medizinischen Fachzeitschriften sowie staatlichen Regulationsbehörden gleichermaßen vernachlässigt und es den Ärzten überlassen, sich ihr eigenes Urteil zu bilden. Die Angaben zu den Determinanten der externen Validität in Studienberichten und systematischen Reviews sind meist unzureichend. Der vorliegende Review befasst sich mit diesen Determinanten, stellt eine Prüfliste für Ärzte vor und gibt Empfehlungen, wie man die externe Validität bei der Planung von RCTs und der Darstellung ihrer Ergebnisse verbessern kann.

„Zwischen den auf RCTs beruhenden Messwerten und dem Nutzen … für die Allgemeinheit klafft ein Abgrund, der immer wieder stark unterschätzt wird.“

(A.L. Cochrane, 1971) [1]

„Bestenfalls zeigt eine Studie, was ein Medikament unter sorgfältiger Beobachtung und bestimmten Einschränkungen bewirken kann. Nicht immer oder zwangsläufig lassen sich dieselben Ergebnisse auch beobachten, wenn dieses Medikament zur allgemeinen Anwendung freigegeben wird.“

(Austin Bradford Hill, 1984) [2]

Randomisierte kontrollierte Studien (RCTs) und systematische Reviews gelten als die zuverlässigsten Methoden zur Bestimmung von Behandlungseffekten. Sie müssen intern valide sein (d. h. Design und Durchführung der Studie müssen die Möglichkeit systematischer Fehler (Bias) auf ein Minimum beschränken) [3, 4]. Um aber klinisch von Nutzen zu sein, muss das Ergebnis auch für eine definierbare Gruppe von Patienten in einem bestimmten klinischen Setting relevant sein; diesen Umstand bezeichnet man gemeinhin als externe Validität, Anwendbarkeit, Übertragbarkeit oder Generalisierbarkeit. Die nützlichen Wirkungen mancher Interventionen wie etwa der Blutdrucksenkung bei unkontrollierter chronischer Hypertonie sind auf die meisten Patienten und Behandlungssituationen übertragbar; die Wirkungen anderer Maßnahmen können dagegen sehr stark von Faktoren wie den Patientencharakteristika, der Applikationsmethode und der Anwendungssituation abhängen. Inwieweit diese Faktoren bei der Planung und Durchführung eines RCT und bei der Darstellung seiner Ergebnisse Berücksichtigung finden, kann einen starken Einfluss auf die externe Validität ausüben.

Die Vernachlässigung der externen Validität in RCTs, systematischen Reviews und Leitlinien wird von Ärzten am häufigsten bemängelt [5–13] und ist eine Erklärung dafür, warum Therapien, die sich in Studien als nützlich erwiesen haben und in Leitlinien empfohlen werden, im Praxisalltag so selten angewendet werden [14–26]. Weder Cochrane noch Bradford Hill waren praktizierende Ärzte; gleichwohl waren ihnen die Grenzen der Methoden, denen sie den Weg bereitet haben, durchaus bewusst. Auch wenn die wenigen systematischen Belege, die uns heute zur Verfügung stehen, bestätigen, dass es RCTs tatsächlich häufig an externer Validität mangelt [27–41], wird dieser Aspekt

heutzutage noch immer von Wissenschaftlern, medizinischen Fachzeitschriften, Studienträgern, Ethikkommissionen, der pharmazeutischen Industrie und den staatlichen Regulationsbehörden gleichermaßen vernachlässigt (Kasten 1) [42–50]. Zugegeben, die Bewertung der externen Validität ist komplex und bedarf eher der ärztlichen als der statistischen Expertise. Sie ist aber essenziell, wenn Therapien im klinischen Alltag bei möglichst vielen Patienten korrekt angewendet werden sollen. Wir können nicht erwarten, dass die Ergebnisse von RCTs und systematischen Reviews für alle Patienten und alle Behandlungssituationen relevant sind (das ist auch nicht mit externer Validität gemeint), doch sollten sie zumindest so angelegt und dargestellt werden, dass Ärzte beurteilen können, auf wen sie sich vernünftigerweise anwenden lassen. In diesem Artikel wird es darum gehen, wie wir die externe Validität bewerten sollten (Kasten 2). Die zur Verdeutlichung ausgewählten Beispiele stammen zwar hauptsächlich aus dem Bereich der Behandlung zerebro- oder kardiovaskulärer Erkrankungen, doch sind die Grundprinzipien für alle Bereiche der Medizin und Chirurgie gleichermaßen gültig.

Kasten 1: Evidenz für die Vernachlässigung der externen Validität in RCTs und systematischen Reviews

- Untersuchungen der internen Validität von RCTs und systematischen Reviews überwiegen die Untersuchungen zur Frage, wie die Ergebnisse am besten in die Praxis umgesetzt werden können, bei weitem [42, 43].
- Die Regeln für die Durchführung von Studien wie etwa die Gute Klinische Praxis [44] erstrecken sich nicht auf Aspekte der externen Validität.
- Zulassungsbehörden wie die US-amerikanische Food and Drug Administration (FDA) verlangen keinen Nachweis, dass ein Medikament einen klinisch nützlichen Therapieeffekt aufweist oder dass eine Studienpopulation für den klinischen Praxisalltag repräsentativ ist [45].
- In den von Kostenträgern wie dem britischen Medical Research Council [46, 47] herausgegebenen Anleitungen für die Planung und Durchführung von RCTs werden Fragen der externen Validität praktisch nicht erwähnt.
- Anleitungen von Ethikkommissionen wie der des britischen Gesundheitsministeriums (Department of Health) [48] weisen darauf hin, dass medizinische Forschungsarbeiten intern valide sein sollten und sprechen einige mit der externen Validität verbundene Fragen an, enthalten aber keine expliziten Empfehlungen im Hinblick darauf, dass die Ergebnisse auch übertragbar sein sollten.
- Leitlinien für die Anfertigung von Studienberichten für RCTs und systematische Reviews befassen sich hauptsächlich mit der internen Validität und schenken Fragen der externen Validität nur sehr wenig Beachtung [49, 50].
- In keiner der Bewertungsskalen für die Beurteilung der Qualität von RCTs wird die externe Validität angemessen berücksichtigt [31].
- Es gibt keine allgemein anerkannten Leitlinien für die Bewertung der externen Validität von RCTs.

Kasten 2: Potenzielle Einflussfaktoren der externen Validität

Studienumgebung
- Gesundheitssystem
- Land
- Rekrutierung von Patienten aus der Primär-, Sekundär- oder Tertiärversorgung
- Auswahl der teilnehmenden Studienzentren
- Auswahl der teilnehmenden Ärzte

Auswahl der Patienten
- Diagnostische und Untersuchungsverfahren vor der Randomisierung
- Eignungs- oder Einschlusskriterien
- Ausschlusskriterien
- Placebo-Run-in-Phase (ohne Behandlung)
- Run-in-Phase mit Behandlung
- „Enrichment"-Strategien
- Verhältnis von randomisierten Patienten zu geeigneten nichtrandomisierten Patienten in den teilnehmenden Zentren
- Anteil der Patienten, die die Randomisierung verweigert haben

Charakteristika der randomisierten Patienten
- Klinische Ausgangscharakteristika
- Ethnische Zugehörigkeit
- Einheitlichkeit der Grunderkrankung
- Stadium im natürlichen Krankheitsverlauf
- Schweregrad der Erkrankung
- Begleiterkrankungen
- Absolute Risiken für einen schlechten Outcome in der Kontrollgruppe

Unterschiede zwischen Studienprotokoll und Praxisalltag
- Studienintervention
- Behandlungszeitpunkt
- Angemessenheit/Relevanz der Kontrollintervention
- Adäquatheit der nicht-studienbezogenen Behandlung – sowohl geplante als auch tatsächliche Behandlung
- Verbot bestimmter nicht-studienbezogener Maßnahmen
- Therapeutische oder diagnostische Fortschritte seit Durchführung der Studie

Zielgrößen und Nachbeobachtung (Follow-up)
- Klinische Relevanz der Surrogatzielgrößen
- Klinische Relevanz, Validität und Reproduzierbarkeit komplexer Skalen
- Effekt der Intervention auf die wichtigsten Komponenten kombinierter Zielgrößen
- Wer misst die Zielgrößen?
- Untersuchung patientenzentrierter Zielgrößen
- Nachbeobachtungshäufigkeit
- Angemessenheit der Nachbeobachtungsdauer

Unerwünschte Therapiewirkungen
- Vollständigkeit der Berichterstattung über relevante unerwünschte Wirkungen
- Häufigkeit von Behandlungsabbrüchen
- Auswahl der Studienzentren und/oder Ärzte auf der Basis ihrer Fertigkeiten oder Erfahrungen
- Ausschluss von Patienten mit Komplikationsrisiken
- Ausschluss von Patienten, bei denen während der Run-in-Phase unerwünschte Wirkungen aufgetreten sind
- Intensität der Studiensicherheitsmaßnahmen

Grenzen der externen Validität

RCTs und systematische Reviews sind die zuverlässigsten Methoden zur Bestimmung mittlerer Behandlungseffekte. Dagegen ist die externe Validität zwangsläufig nicht vollkommen gege-

ben – zumindest theoretisch –, denn das Ziel ist nicht die Messung des Behandlungsnutzens im klinischen Alltag. Das Ansprechen auf und/oder die Compliance mit einer Behandlung kann stark vom Arzt-Patient-Verhältnis [51–53], Placeboeffekten [54, 55] und den Präferenzen des Patienten beeinflusst sein [56–58]. Doch versuchen Prüfärzte zu Recht, die Einflüsse dieser Faktoren möglichst durch verblindete Behandlungszuteilung, Placebokontrollen und den Ausschluss von Patienten oder Ärzten, die starke Therapiepräferenzen erkennen lassen, zu eliminieren. Solche Maßnahmen erhöhen die interne Validität eines RCT, führen im klinischen Alltag oft jedoch – insbesondere bei patientenzentrierten Zielgrößen – zu einer Unterschätzung des Behandlungsnutzens.

Patientenpräferenzen können im Hinblick auf die externe Validität ein besonderes Problem darstellen. So zeigen etwa einige Frauen mit einem Mammakarzinom im Frühstadium eine starke Präferenz für eine Lumpektomie, wohingegen andere mit dem Gedanken, dass die gesamte Krebsgeschwulst durch eine Mastektomie entfernt wird, weitaus glücklicher sind. Nur Frauen ohne eine starke Präferenz für eine bestimmte Behandlung konnten für die relevanten RCTs rekrutiert werden, und gerade einmal 10% stimmten einer randomisierten Behandlungszuteilung zu [59]. Wenn RCTs für eine Behandlung einen größeren Vorteil ergeben, dann stellt die externe Validität kein Problem dar. Schwierigkeiten entstehen dann, wenn eine Behandlung nur mäßig wirksamer ist, der Patient aber eine starke persönliche Präferenz für die weniger wirksame Option erkennen lässt. Wären die Ergebnisse der RCTs über Mammaoperationen, vor allem im Hinblick auf das psychische Wohlergehen, dieselben gewesen, wenn die Patientinnen der Behandlung randomisiert zugeteilt worden wären?

Diese unvermeidlichen Einschränkungen stellen die Ergebnisse von RCTs und systematischen Reviews zwar nicht in Frage – und sie werden hier auch nur der Vollständigkeit halber erwähnt –, doch sollte die Bedeutung von Patientenpräferenzen, Placeboeffekten und Arzt-Patient-Verhältnis außerhalb von Studien nicht unterschätzt werden. Dieser Umstand lässt sich vielleicht am besten anhand der Popularität alternativer Behandlungsansätze wie der Homöopathie veranschaulichen, bei denen solche Faktoren die einzig wirksamen Inhaltsstoffe darstellen. Im verbleibenden Teil dieses Reviews werden wir uns auf diejenigen Faktoren der Planung von RCTs und systematischen Reviews und ihrer Ergebnisdarstellung konzentrieren, die die externe Validität einschränken, aber nicht unvermeidlich sind.

Studienumgebung

Häufig werden Bedenken geäußert, dass die in der Sekundär- oder Tertiärversorgung durchgeführten Studien nicht auf die Primärversorgungsbereiche übertragbar sind [26–29], doch die Studienumgebung kann auch noch auf verschiedene andere Weisen die externe Validität beeinflussen.

Gesundheitssystem

Auch Unterschiede zwischen verschiedenen Gesundheitssystemen können sich auf die externe Validität auswirken. So wurden beim European Carotid Surgery Trial (ECST) [60], einem RCT zur Endarteriektomie bei kurz zurückliegender symptomatischer Karotisstenose, nationale Unterschiede hinsichtlich der Geschwindigkeit festgestellt, mit der Patienten untersucht wurden. Dabei ergab sich eine mediane Verzögerung zwischen dem Zeitpunkt der letzten Symptome bis zur Randomisierung von mehr als zwei Monaten in Großbritannien im Vergleich zu drei Wochen in Belgien und den Niederlanden. Abbildung 1 zeigt, dass getrennt durchgeführte Studien in diesen Gesundheitssystemen wegen des engen Zeitfensters der Schlaganfallprophylaxe zu ganz anderen Ergebnissen geführt hätten [61]. Auf diese Unterschiede wurde in keiner der ECST-Publikationen oder den da-

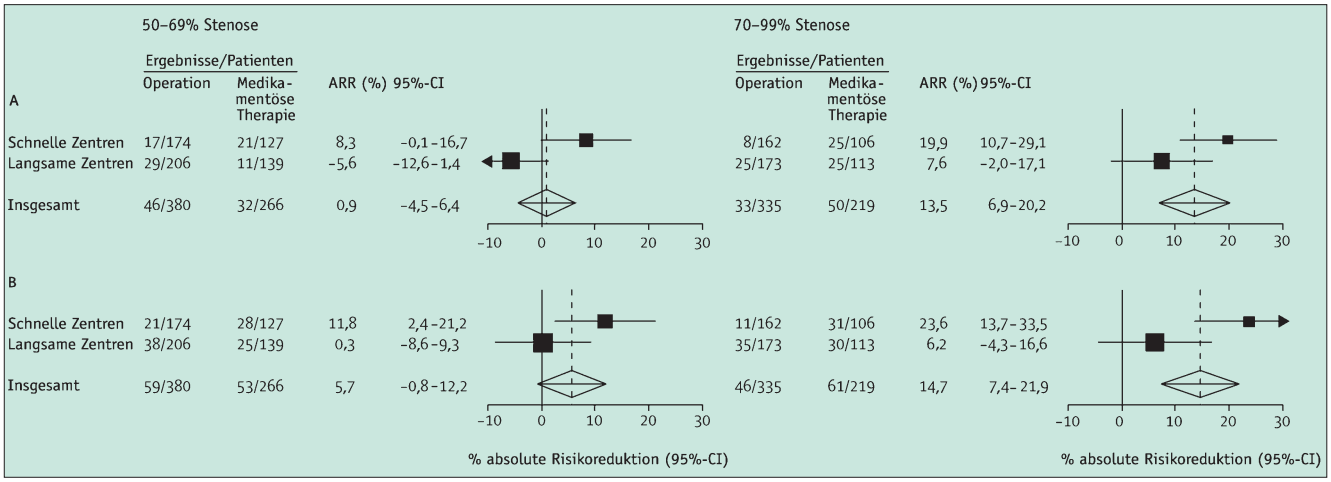

Abb. 1. Absolute Reduktion (ARR) des 5-Jahres-Risikos für ipsilateralen ischämischen Schlaganfall (obere Zeilen) und für alle Schlaganfälle oder Tod (untere Zeilen) bei Operation in ECST-Zentren [60].
Die mediane Verzögerung vom letzten symptomatischen Ereignis bis zur Randomisierung betrug für schnelle Zentren maximal 50 Tage (schnelle Zentren) im Vergleich zu langsamen Zentren (> 50 Tage). Die Daten sind getrennt nach Patienten mit mittelschwerer (50–69%, linke Seite der Tabelle) und schwerer (70–99%, rechte Seite der Tabelle) Karotisstenose dargestellt. **A**: ipsilateraler ischämischer Schlaganfall oder perioperativer Schlaganfall/Tod. **B**: jeder Schlaganfall oder perioperativer Tod

raus abgeleiteten Leitlinien aufmerksam gemacht. Ähnliche Unterschiede hinsichtlich der Leistungen der verschiedenen Gesundheitssysteme bestehen sicher auch bei anderen Erkrankungen, und letztlich ist natürlich auch noch die allgemeinere Frage offen, inwieweit die Ergebnisse der in Industrienationen durchgeführten Studien auf Länder der Dritten Welt übertragbar sind.

Land

Selbst wenn sich die Gesundheitssysteme ähnlich sind, können noch andere nationale Unterschiede die Übertragbarkeit von Studienergebnissen beeinflussen. Kommen wir noch einmal auf unser Beispiel der zerebrovaskulären Erkrankungen zurück. Zwischen den einzelnen Ländern bestehen neben zahlreichen Unterschieden in den diagnostischen und Behandlungsmethoden [62] auch wichtige ethnische Unterschiede bezüglich Krankheitsanfälligkeit und natürlichem Krankheitsverlauf [63], die allesamt Einfluss auf die externe Validität von Studienergebnissen nehmen können. Auch aus anderen Bereichen der Medizin lassen sich Beispiele anführen, darunter etwa die beträchtliche Heterogenität zwischen den Studien über BCG zur Tuberkuloseprävention in verschiedenen Populationen, bei denen sich eine hohe Wirksamkeit in den nördlichen Ländern ergab, die jedoch allmählich abnahm ($p < 0{,}00001$), je südlicher die Lage des Landes war [64]. Ferner sind auffällige nationale Unterschiede in der Anwendung zusätzlicher Behandlungen, also Therapien außerhalb von Studien, erkennbar. In einem internationalen RCT über die Anwendung von Aspirin und Heparin bei nahezu 20.000 Patienten mit akutem ischämischem Schlaganfall wurde in Italien Glycerin bei 50% der 1473 Patienten angewendet im Vergleich zu 3% anderswo, Steroide bei 32% der 225 Patienten in der Türkei im Gegensatz zu 4% anderswo und Hämodilution bei 44% der 597 Patienten in Österreich und der Tschechischen Republik verglichen mit 3% anderswo [65]. Noch gravierendere Länderunterschiede wurden bei der Anwendung zweier wichtiger, nicht in der Studie untersuchter Operationsverfahren in der ECST erfasst (Abb. 2). Die Evidenz lässt darauf schließen, dass die Anwendung beider Verfahren das Operationsrisiko beeinflusst, aber ungeachtet dieser Tatsache zeigen die Daten, wie sehr die klinische Praxis in den einzelnen Ländern variiert. Die in einem Land durchgeführten RCTs lassen sich meist zwar auf andere Länder übertragen, die Übertragbarkeit der Ergebnisse sollte aber nicht als selbstverständlich betrachtet werden.

Auswahl der teilnehmenden Zentren und Ärzte

Die Auswahl der teilnehmenden Zentren aus der Sekundär- im Gegensatz zur Primärversorgung hat eindeutig Auswirkungen auf die externe Validität, kann aber auch RCTs über Interventionen, die sich auf die Sekundärversorgung beschränken, unterminieren, wenn diese auf Spezialeinrichtungen begrenzt bleiben [66–68]. In einem systematischen Review über die laparoskopische Cholezystektomie ergab sich beispielsweise, dass alle 15 RCTs ausschließlich in Universitätskrankenhäusern stattgefunden hatten [67]. Probleme entstehen auch, wenn die teilnehmenden Ärzte aufgrund ihrer Erfolgsbilanz ausgewählt werden. Die Ergebnisse der ACAS- (Asymptomatic Carotid Artery Surgery) Studie haben etwa gezeigt, dass die Endarteriektomie bei asymptomatischer Karotisstenose das absolute 5-Jahres-Risiko für Schlaganfall um ca. 5% senkte [69]. In die ACAS-Studie wurden jedoch nur Chirurgen mit einem hohen Sicherheitsstandard aufgenommen: 40% der Bewerber wurden gleich zu Beginn der Studie abgelehnt [70] und später diejenigen von der weiteren Teilnahme ausgeschlossen, die im Rahmen der Studie unerwünschte Operationsergebnisse erzielt hatten. Die Vorteile der Operation, die sich in der ACAS-Studie ge-

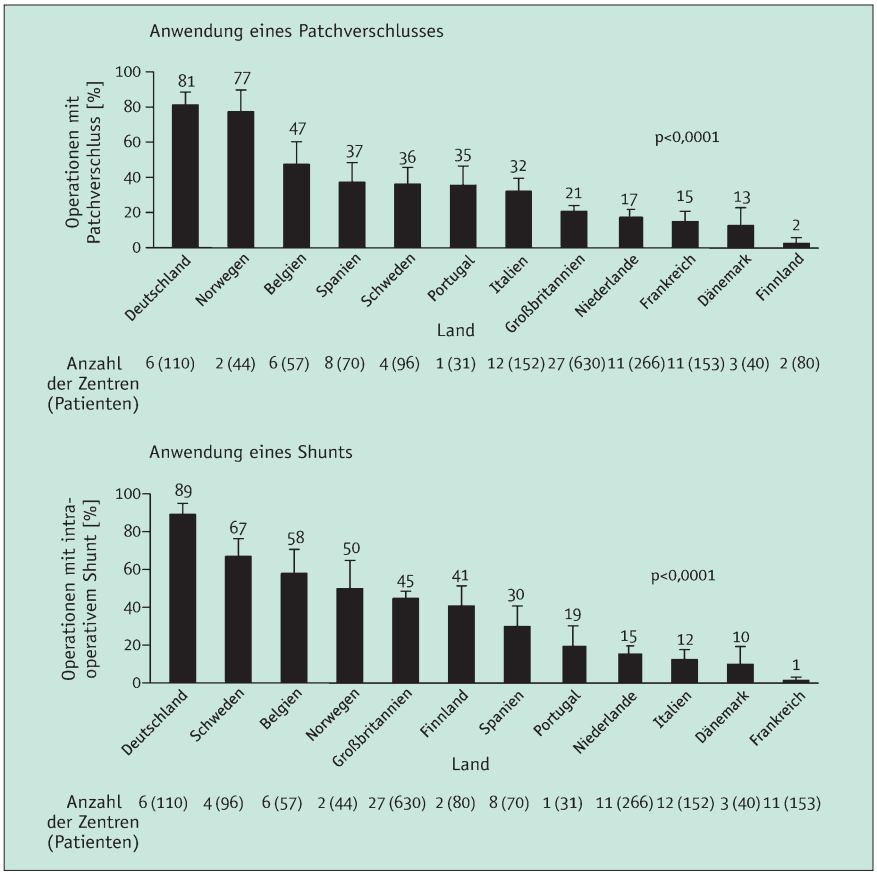

Abb. 2. Schwankungen in der Häufigkeit der Anwendung von zwei zusätzlichen operativen Verfahren bei Karotisendarteriektomie [Patchangioplastie (oberes Balkendiagramm) und intraoperativer Shunt (unteres Balkendiagramm)] im Rahmen der ECST [60], nach Ländern geordnet

Z. ärztl. Fortbild. Qual. Gesundh.wes. (2006) 100; 303–317
http://www.elsevier.de/zaefq

zeigt hatten, waren zum großen Teil also auf das niedrige Operationsrisiko zurückzuführen [69]. Abbildung 3 vergleicht die ACAS-Risiken mit den Ergebnissen einer Meta-Analyse von 46 chirurgischen Fallserien, in denen die Operationsrisiken während der fünf Jahre nach der ACAS-Studie veröffentlicht wurden [71]. Die perioperative Mortalität war achtmal höher als in der ACAS-Studie (1,11% vs. 0,14%; $p = 0{,}01$) und das Risiko für Schlaganfall und Tod in vergleichbaren Studien, in denen das Ergebnis durch einen Neurologen bewertet wurde, ca. dreimal höher (4,30% vs. 1,50%; $p < 0{,}001$). Nun sollten in Studien keine Zentren eingeschlossen werden, die nicht in der Lage sind, die sichere Behandlung der Patienten zu gewährleisten. Doch sollte die Auswahl der Zentren auch nicht so anspruchsvoll sein, dass sich die Ergebnisse nicht mehr auf den Klinikalltag übertragen lassen. So ist es eher unwahrscheinlich, dass die aus der ACAS-Studie ausgeschlossenen Chirurgen außerhalb des Studienrahmens plötzlich aufgehört haben zu operieren.

Auswahl der Patienten

Nur ein geringer Anteil aller Patienten mit einer spezifischen Erkrankung nimmt an einer bestimmten Studie teil. So wurde z. B. nur etwa einer von 200 bis 300 Patienten, die sich in Nordamerika damals einer Karotisendarteriektomie unterzogen, in die großen multizentrischen RCTs [72] aufgenommen; ähnliche Raten werden für die Mammakarzinomstudien angegeben [73]. Diese niedrigen Raten bedeuten, dass die Rekrutierung für solche Studien viele Jahre dauert, für die externe Validität aber kein Problem darstellen, solange die in den teilnehmenden Zentren randomisierten Patienten für die Grundgesamtheit repräsentativ sind. Wie unten zu zeigen sein wird, ist dies nicht immer der Fall.

Selektion vor der Prüfung der Ein- und Ausschlusskriterien

Nicht selten lösen hochselektive Studieneinschlusskriterien Bedenken aus, doch es gibt noch verschiedene frühere Stadien im Auswahlprozess, die viel problematischer sein können. Abbildung 4 zeigt, dass in der lokalen Bevölkerung, die von einem an einer Studie teilnehmenden Zentrum versorgt wird, der Anteil der Patienten mit einer bestimmten Erkrankung, die für den Studieneinschluss in Betracht gezogen werden, oftmals deutlich unter 1% liegt. Sehen wir uns dazu beispielsweise eine Studie über ein neues blutdrucksenkendes Medikament an, die wie die Mehrzahl ähnlicher Studien in einem Krankenhaus durchgeführt wird. Weniger als 10% der Hypertoniepatienten werden in Krankenhäusern behandelt, und diese Gruppe unterscheidet sich von den Hypertoniepatienten in der Primärversorgung. Darüber hinaus nimmt nur einer von zehn Ärzten, die in diesem Krankenhaus Hochdruckpatienten behandeln, auch an dieser Studie teil, und dieser besondere Arzt hat es hauptsächlich mit Querüberweisungen junger Patienten mit therapierefraktärer Hypertonie zu tun. Noch bevor also überhaupt über Ein- oder Ausschlusskriterien nachgedacht wird, stellen die potenziellen Kandidaten für eine solche Studie schon gar keine repräsentative Stichprobe der Patienten aus der Lokalbevölkerung mehr dar.

Selektion nach Ein- und Ausschlusskriterien

Anschließend werden die Patienten weiter anhand der Studieneinschlusskriterien selektiert (Abb. 4). Manche RCTs schließen Frauen und viele Studien auch ältere Menschen aus [74, 75]. In einem Review von 214 Medikamentenstudien zu akutem Myokardinfarkt wurde festgestellt, dass mehr als 60% der Studien Patienten über 75 Jahren ausgeschlossen hatten [74]. Viele RCTs schließen ferner Patienten mit häufig vorkommenden Begleiterkrankungen aus. Studien über Thrombozytenaggregationshemmer oder nichtsteroidale Antiphlogistika (NSAIDs) etwa schließen häufig die Rekrutierung von Patienten mit einer Dyspepsievorgeschichte aus, auch wenn davon im Praxisalltag mehr als 50% der älteren Patienten betroffen sind [76, 77]. Die Ausschlussraten können sehr hoch sein. Bei akutem Schlaganfall ergab eine Studie, dass von dem geringen Anteil der Patienten, die schnell genug zugelassen wurden, um für eine Thrombolyse in Frage zu kommen [78], 96% auf der Grundlage der verschiedenen anderen Kriterien des fraglichen RCT gar nicht geeignet gewesen wären [79]. In einer weiteren Studie über akuten Schlaganfall musste ein Studienzentrum über zwei Jahre 192 Patienten screenen, um *einen* geeigneten Patienten zu finden [80]. Ein Review von 41 RCTs der US National Institutes of Health ermittelte eine durchschnittliche Ausschlussrate von 73% [39].

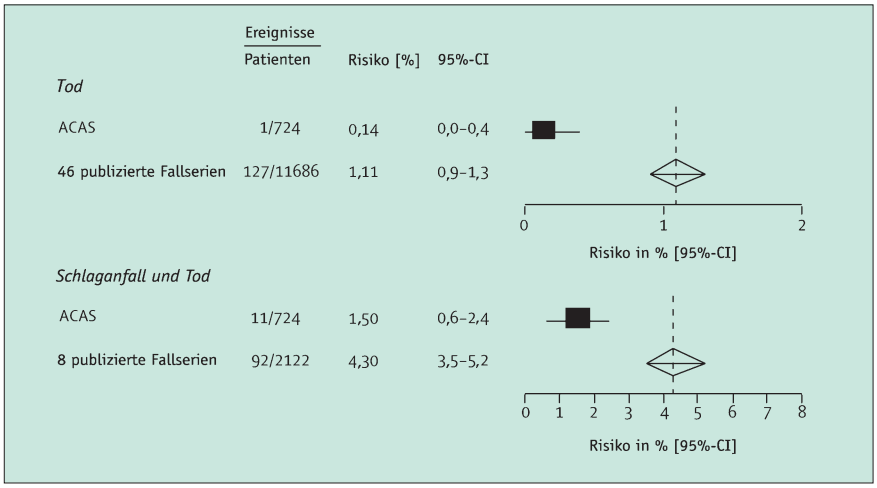

Abb. 3. Gesamtergebnisse einer Meta-Analyse aller zwischen 1990 und Ende 2000 veröffentlichten Studien zum Risiko bei Karotisendarteriektomie wegen asymptomatischer Stenosen [71], in denen das Operationsrisiko für Tod (obere Zeilen) und Schlaganfall und Tod (untere Zeilen) mit den Daten für dieselben Risiken aus der ACAS-Studie [69] verglichen wurde

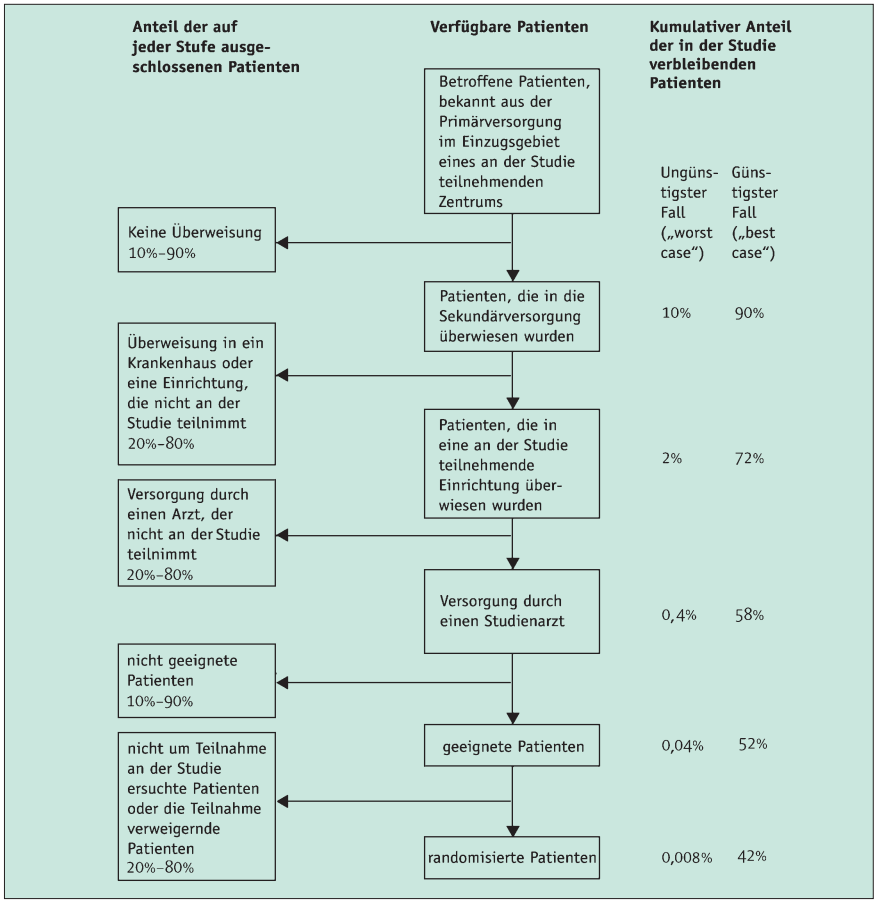

Abb. 4. Schaubild zur Veranschaulichung der Auswirkungen einer am für den Klinikalltag orientierten typischen mehrstufigen Selektion auf den Anteil der Patienten im Einzugsbereich eines Studienzentrums, die für einen in der Sekundärversorgung durchgeführten RCT rekrutiert werden.

Selektion über die Ein- und Ausschlusskriterien hinaus

Strenge Einschlusskriterien können die externe Validität von RCTs beeinträchtigen und im Praxisalltag zu niedrigeren Behandlungsraten führen [81], doch immerhin sind diese Kriterien für eine Überprüfung auch verfügbar oder sollten es zumindest sein (siehe unten). Schwieriger ist es, wenn Studien mit scheinbar vernünftigen Einschlusskriterien am Ende mit hoch selektierten Studienpopulationen dastehen. Eine Rekrutierung von weniger als 10% der Patienten mit der betreffenden Erkrankung ist in den an pragmatischen RCTs teilnehmenden Zentren in allen Bereichen von Medizin und Chirurgie, für die Daten erhoben wurden, gang und gäbe [24, 33, 82–87]. Diese niedrigen Rekrutierungsraten sind zum Teil darauf zurückzuführen, dass die teilnehmenden Ärzte eine über die Einschlusskriterien hinausgehende zusätzliche Selektion vornehmen. Die für RCTs rekrutierten Patienten unterscheiden sich von den geeigneten, aber nicht rekrutierten Patienten im Hinblick auf Alter, Geschlecht, ethnische Zugehörigkeit, Krankheitsschweregrad, Bildungsstand, sozialen Status und Wohnort [30, 34, 35, 37, 59, 74, 75, 88–90]. Auch die Prognosen der Patienten, die in RCTs aufgenommen werden, sind meist besser als derjenigen, die nicht an Studien teilnehmen [91], oftmals sogar sehr viel besser [6, 92]. Und doch ist eine solche hochselektive Rekrutierung nicht unvermeidlich. Für die GISSI-1-Studie zur Thrombolyse bei akutem Myokardinfarkt wurden z. B. 90% der Patienten rekrutiert, die innerhalb von 12 h nach dem Indexereignis mit einer endgültigen Diagnose und ohne Kontraindikationen stationär aufgenommen wurden [93]. Damit verfügt diese Studie über eine exzellente externe Validität und gehört zu den sehr wenigen RCTs über akuten Myokardinfarkt, in denen die Mortalität in der Kontrollgruppe (13%) wenigstens entfernt mit den Werten im damaligen klinischen Alltag übereinstimmte.

Run-in-Phasen

Um Patienten für eine Studie zu selektieren oder auszuschließen, greift man vor der Randomisierung häufig auf Run-in-Phasen zurück [94]. In einer Placebo-Run-in-Phase erhalten alle geeigneten Patienten Placebo, und Patienten mit mangelnder Compliance werden ausgeschlossen [95]. Für eine solche Entscheidung kann es zwar gute Gründe geben, doch hohe Ausschlussraten mindern die externe Validität. In einer Studie über die Wirkungen von Salzen auf den Blutdruck wurden beispielsweise 93% der Patienten in der Placebo-Run-in-Phase ausgeschlossen [96]. Sehr viel höher ist allerdings die Wahrscheinlichkeit für eine starke Beeinträchtigung der externen Validität in Run-in-Phasen mit aktiver Behandlung, in denen Patienten ausgeschlossen werden, wenn unerwünschte Ereignisse auftreten oder sie Anzeichen dafür erkennen lassen, dass die Behandlung bei ihnen möglicherweise nicht wirkt. So wurden z. B. in zwei RCTs über Carvedilol, einen vasodilatatorisch wirkenden β-Blocker, zur Behandlung der chronischen Herzinsuffizienz 6% und 9% der geeigneten Patienten in den Run-in-Phasen mit Behandlung [97, 98] wegen einer Verschlechterung der Herzinsuffizienz und anderer unerwünschter Ereignisse, von denen einige tödlich verliefen, ausgeschlossen. In beiden Studien waren die Komplikationsraten in der nachfolgenden Phase (nach der Randomisierung) deutlich geringer als in der Run-in-Phase [97, 98].

Anreicherungsstrategien

Patienten, die wahrscheinlich gut auf die Behandlung ansprechen, werden gelegentlich aktiv rekrutiert [99–101]. In einigen Studien über antipsychotische Medikamente wurden die Patienten, die zuvor gut auf solche Präparate angesprochen hatten, selektiv rekrutiert [102]. In anderen Studien wurden

Patienten, die in einer Run-in-Phase kein Ansprechen gezeigt hatten, ausgeschlossen. In einem RCT über den Acetylcholinesterasehemmer Tacrin zur Behandlung des Morbus Alzheimer wurden 632 Patienten für eine 6-wöchige Präselektionsphase (sog. „Enrichment", Anreicherung) rekrutiert, in der die Patienten unterschiedliche Tacrin-Dosierungen oder Placebo erhielten [103]. Nach einer Auswaschphase wurden nur die 215 Patienten (34%), bei denen sich in der Präselektionsphase unter Tacrin eine messbare Besserung gezeigt hatte, in der Hauptphase der Studie für die Behandlung mit Tacrin (in der für sie wirksamsten Dosis) versus Placebo randomisiert. In diesem Fall war die externe Validität eindeutig beeinträchtigt.

Angaben zur Patientenselektion

Die Anzahl der geeigneten, nichtrandomisierten Patienten kann dokumentiert werden, sie lässt sich aber nur schwer zuverlässig bestimmen und unterschätzt die Selektion, da Logbücher meist nur die zum Studienarzt überwiesenen Patienten erfassen. Ein weiterer nützlicher Index ist die Anzahl der Patienten, die zur Teilnahme aufgefordert werden und diese verweigern, aber normalerweise wird keiner der beiden Werte angegeben. Ein deutlich größeres Hindernis für die Bewertung der externen Validität stellen jedoch unzulängliche Angaben zu den Studieneinschlusskriterien dar [104]. Die CONSORT-Leitlinien [49] und die Allgemeinen Anforderungen an Manuskripte zur Publikation in biomedizinischen Fachzeitschriften (Uniform Requirements for Manuscripts Submitted to Biomedical Journals) (http://www.ICMJE.org) verlangen, dass alle Auswahlkriterien angegeben werden, doch ein Review von Studien, die die US National Institutes of Health zur Veröffentlichung von Warnhinweisen veranlassten, ergab, dass von den durchschnittlich 31 Auswahlkriterien nur 63% im Hauptstudienbericht und nur 19% in den Warnhinweisen für Ärzte dargelegt wurden [105]. Die unzureichende Ergebnisdarstellung ist ein relevantes Problem von Sekundärveröffentlichungen wie systematischen Reviews und Behandlungsleitlinien, in denen aus Platzgründen und zugunsten eines knappen Berichtstils meist auf detaillierte Angaben zu den Ein- und Ausschlusskriterien von Studien oder anderen Determinanten der externen Validität verzichtet wird. Dasselbe gilt auch für das Marketing von Pharmaunternehmen, obwohl hier andere Gründe für das Verschleiern einer mangelhaften externen Validität vorliegen mögen.

Charakteristika der randomisierten Patienten

Studienberichte enthalten gewöhnlich die Ausgangscharakteristika der rando-

Tabelle 1. Ausgangscharakteristika und Outcomes der für die ECST randomisierten Patienten [60].

	Operation (n = 1807)		Keine Operation (n = 1211)
	Nicht operiert (n = 62)	Operiert (n = 745)	
Demographische Daten			
Männliches Geschlecht	36 (58%)	1263 (72%)	869 (72%)
Alter (in Jahren)	64,1 (8,7)	62,5 (8,1)	62,3 (8,0)
Zerebrovaskuläre Ereignisse in den vergangenen 6 Monaten			
TIA oder Schlaganfall, halbseitig	56 (90%)	1495 (85%)	1038 (86%)
Nur okuläre Ereignisse	6 (10%)	250 (15%)	173 (14%)
Bleibende neurologische Symptome	2 (3%)	106 (6%)	78 (7%)
Tage seit den letzten Symptomen	74 (56)	62 (53)	62 (52)
Andere klinische Daten			
Früherer Schlaganfall	2 (3%)	101 (6%)	78 (7%)
Systolischer Blutdruck (mmHg)	154 (27,2)	151 (22,3)	150,2 (21,3)
Diastolischer Blutdruck (mmHg)	89,0 (13,0)	86,2 (11,4)	86,3 (10,8)
Angina pectoris	15 (24%)	305 (18%)	190 (16%)
Früherer Myokardinfarkt	7 (11%)	219 (13%)	136 (11%)
Frühere Koronararterien-OP	2 (3%)	47 (3%)	23 (2%)
Periphere Gefäßerkrankung	7 (11%)	292 (16%)	203 (17%)
Diabetes	8 (13%)	208 (12%)	145 (12%)
Derzeitiges Zigarettenrauchen	25 (40%)	844 (48%)	557 (46%)
Cholesterinspiegel im Blut (mmol/l)	6,4 (1,6)	6,4 (1,4)	6,4 (1,4)
Mittelwert symptomatischer Karotisstenosen	60% (25)	62% (21)	59% (22)
Mittelwert kontralateraler Karotisstenosen	37% (26)	42% (26)	37% (27)

Patienten, die für die Operation randomisiert waren, aber nicht operiert wurden, im Vergleich zu operierten Patienten und zu Patienten, die zur alleinigen medikamentösen Therapie randomisiert wurden.
TIA = transiente ischämische Attacke
Die Daten werden als Mittelwerte (Standardabweichung) oder Anzahl (%) angegeben.

misierten Patienten, damit – so das Argument – Ärzte die externe Validität durch Vergleich mit ihren eigenen Patienten bewerten können [49, 50]. Diese Theorie klingt vernünftig, doch können die Ausgangscharakteristika irreführend sein. Wie schwierig es ist, die Ausgangscharakteristika zu extrapolieren, zeigt sich an den Patienten, die in der ECST-Studie für die Endarteriektomie randomisiert worden waren [60], aber nicht operiert wurden, weil ihr Operateur und/oder Anästhesist sie für zu gebrechlich hielt. Auch wenn sich dieser klinische Eindruck durch einen im Vergleich zu den Patienten, die für die medikamentöse Behandlung randomisiert wurden, deutlich schlechteren Outcome während der Nachbeobachtung bestätigte (5-Jahres-Risiken: Schlaganfall 36% vs. 18%, $p < 0{,}001$; Schlaganfall oder Tod 52% vs. 27%; $p < 0{,}0001$), hatten sich ihre Ausgangscharakteristika nicht unterschieden (Tabelle 1).

Ein Patient kann sich auch in scheinbar irrelevanter Weise von der Studienpopulation unterscheiden, was aber relevante Auswirkungen auf die externe Validität haben kann. Tabelle 2 zeigt z. B. die Ausgangscharakteristika der Patienten, die in zwei RCTs zur sekundären Schlaganfallprävention für die Behandlung mit Warfarin randomisiert wurden [106–108]. In einer Studie handelte es sich um Patienten mit Vorhofflimmern und in der andern um Patienten mit Sinusrhythmus. Man könnte erwarten, dass dieser Unterschied einen Einfluss auf das Risiko für einen weiteren ischämischen Schlaganfall hat, nicht aber auf die Sicherheit von Warfarin. Tatsächlich war aber nach Adjustierung nach den Ausgangscharakteristika und der Intensität der Antikoagulation das intrakranielle Blutungsrisiko unter Warfarin in der SPIRIT-Studie (Stroke Prevention in Reversible Ischaemia Trial) 19mal so hoch ($p < 0{,}0001$) wie in der EAFT-Studie (European Atrial Fibrillation Trial) (Tabelle 2) [108]. Scheinbar irrelevante Unterschiede zwischen den Patienten können relevante Auswirkungen auf Risiken und Nutzen von Therapien haben. Noch deutlicher wird diese Tatsache in der SPIRIT-Studie durch das Beispiel der Patienten, bei denen zu Beginn der Studie mittels Brain-Imaging-Verfahren eine Leukoaraiose nachgewiesen wurde und die unter Warfarin ein 9mal höheres intrakranielles Blutungsrisiko aufwiesen als Patienten ohne Leukoaraiose [106, 108].

Es gibt noch viele andere mit Patientencharakteristika verbundene Faktoren, die die Relevanz einer Studie für einen bestimmten Patienten bestimmen können, darunter auch die zugrunde liegende Pathologie, der Krankheitsschweregrad, das Krankheitsstadium im natürlichen Krankheitsverlauf, Begleiterkrankungen und das wahrscheinliche absolute Risiko für einen ohne Therapie ungünstig verlaufenden Outcome.

Tabelle 2. Ausgangscharakteristika und Outcomes von Patienten, die in der EAFT- [106] und SPIRIT- [107] Studie für die Antikoagulation mit Warfarin randomisiert wurden.

	SPIRIT (n = 651)	EAFT (n = 225)
Ausgangscharakteristika		
Männliches Geschlecht	66%	55%
Alter > 65 Jahre	47%	81%
Hypertonie	39%	48%
Angina pectoris	9%	11%
Myokardinfarkt	9%	7%
Diabetes	11%	12%
Leukoaraiose im Hirn-CT	7%	14%
Outcomes während der Studie		
INR während der Studie (Mittelwert, Standardabweichung)	3,3 (1,1)	2,9 (0,7)
Patientenjahre Nachbeobachtung	735	507
Intrakranielle Blutungen	27	0*
Extrakranielle Blutungen	26	13
Adjustierte Hazard-Ratio (95%-CI)*		
Intrakranielle Blutungen	19,0 (2,4 bis 250) $p < 0{,}0001$	
Extrakranielle Blutungen	1,9 (0,8 bis 4,7) $p = 0{,}15$	

INR = International Normalised Ratio (Thromboplastinzeit); CT = Computertomographie.
*Intrakranielle Blutungen wurden zwar nicht nachgewiesen, bei zwei Schlaganfällen wurde aber kein CT aufgenommen. Bei der Berechnung der adjustierten Hazard-Ratio für Hämorrhagien wurden diese beiden Schlaganfälle als durch intrakranielle Blutungen bedingt gewertet.

Intervention, Kontrollbehandlung und Behandlung vor bzw. außerhalb der Studie

Die externe Validität kann auch beeinträchtigt sein, wenn Studien Behandlungsprotokolle untersuchen, die von der üblichen klinischen Praxis abweichen. Vor der Randomisierung für die RCTs über Endarteriektomie bei symptomatischer Karotisstenose mussten die Patienten z. B. von einem Neurologen diagnostiziert werden und sich einer konventionellen arteriellen Angiographie unterziehen [109], was beides in vielen Zentren nicht zur Routine gehört. Die Studienintervention selbst kann ebenfalls von der üblichen Praxis abweichen, wie etwa die Zubereitung und Bioverfügbarkeit eines Medikaments oder die Art des bei einer Operation verwendeten Anästhetikums. Dasselbe kann auch auf die Behandlung in der Kontrollgruppe einer Studie zutreffen, in der eine besonders niedrige Dosis des Vergleichspräparates angewendet wird oder die auf andere Weise gegenüber der besten derzeitigen Vorgehensweise (Best Current Practice) abfällt. Möglicherweise wird die externe Validität auch durch zu strenge Einschränkungen bei der Anwendung von nicht-studienbezogenen Therapien unterlaufen. So sind etwa Antihypertensiva oder Medikamente zur Behandlung der Herzinsuffizienz bei älteren Patienten außerhalb der Studienumgebung, die ihre NSAIDs nicht absetzen

können, weniger wirksam [110]. Alle Verbote von nicht-studienbezogenen Therapien sollten neben den Einzelheiten zu relevanten, außerhalb der Studie angewandten Behandlungen in der Hauptpublikation der Studie dargelegt werden. Auch der Zeitpunkt vieler Interventionen kann sich, wie in Abbildung 1 am Beispiel der Endarteriektomie bei rezenter symptomatischer Karotisstenose gezeigt, als erfolgskritisch erweisen und sollte, wenn er relevant ist, auch angegeben werden.

Zielgrößen und Nachbeobachtung

Die externe Validität eines RCT richtet sich auch danach, ob die Zielgrößen (auch Endpunkte genannt) klinisch relevant waren. Dies kann von so subtilen Fragen abhängen, wer den Endpunkt tatsächlich untersucht hat, wie sich am Beispiel der geringeren operativen Risiken der Endarteriektomie in den Studien zeigen lässt, in denen die Patienten statt durch Neurologen von Chirurgen beurteilt wurden [111], ist jedoch meist davon abhängig, was wann gemessen wurde.

Surrogatzielgrößen

Viele Studien verwenden Surrogatzielgrößen, also im Allgemeinen biologische oder bildgebende Marker, die als indirekte und damit Ersatzmessgrößen für den Effekt der Behandlung auf den klinischen Endpunkt gelten. Surrogatzielgrößen sind jedoch häufig nicht nur von fragwürdiger klinischer Relevanz, sondern oftmals sogar irreführend. Alle Therapien in Tabelle 3 hatten eine relevante nützliche Wirkung auf eine Surrogatzielgröße gezeigt, doch obwohl jede Surrogatzielgröße in Beobachtungsstudien mit einem relevanten klinischen Endpunkt korreliert war, erwiesen sich die Therapien in nachfolgend durchgeführten großen RCTs, in denen diese klinischen Endpunkte untersucht wurden, als unwirksam oder sogar schädlich [112–124].

Skalen

Zuweilen werden in RCTs komplexe Skalen eingesetzt, die häufig aus willkürlich kombinierten Symptomen und klinischen Anzeichen bestehen [125, 126]. So wurden etwa in einem Review von 196 RCTs über NSAIDs zur Behandlung der rheumatoiden Arthritis mehr als 70 verschiedene Outcome-Skalen gefunden [127] und in einem Review von 2000 RCTs zur Schizophrenie sogar 640 Skalen, von denen viele jeweils für die spezifische Studie entwickelt worden waren und deren Validität oder Reliabilität von keinerlei Daten gestützt wurden [128]. Die Wahrscheinlichkeit, dass mit diesen nicht validierten Skalen signifikante Therapieeffekte auftraten, war höher als bei anerkannten Skalen [129]. Außerdem lässt sich die klinische Bedeutung scheinbarer Behandlungseffekte (z. B. eine mittlere Reduktion von 2,7 Punkten auf einer aus verschiedenen Symptomen und Anzeichen bestehenden 100-Punkte-Outcome-Skala) meist nur schwer einordnen.

Patientenzentrierte Endpunkte

Einfache klinische Endpunkte haben meist die höchste externe Validität. Dies gilt aber nur dann, wenn sie auch die Prioritäten der Patienten widerspiegeln. Abbildung 5 beispielsweise zeigt die Ergebnisse einer Studie, in der Patienten mit Multipler Sklerose und ihre Ärzte unabhängig voneinander gebeten wurden, drei Krankheitsaspekte mit den stärksten Auswirkungen auf die Lebensqualität zu benennen [130]. Bei den Ärzten lag der Schwerpunkt hauptsächlich auf den körperlichen Effekten der Erkrankung, während die Sorgen der Patienten eher um psychische Gesundheit, emotionales Wohlbefinden, Allgemeinzustand und Vitalität kreisten, die in RCTs häufig nicht untersucht werden. Wichtig ist auch, die klinischen Endpunkte so zu formulieren, dass sie für die Patienten möglichst relevant sind, auch wenn solche Schätzungen die statistische Power (Trennschärfe) der Studie verringern. So interessieren sich z. B. Patienten mit Epilepsie sehr viel mehr für den Anteil der Patienten, die in RCTs über Antikonvulsiva anfallsfrei werden, als für Änderungen der mittleren Anfallshäufigkeit [131, 132].

Kombinierte Endpunkte

In vielen Studien werden verschiedene Ereignisse in der primären Zielgröße zusammengefasst. Eine solche Kombi-

Tabelle 3. Beispiele für Studien mit irreführenden Surrogatzielgrößen.

	Erkrankung	Surrogatzielgröße	Klinischer Endpunkt
Therapie			
Fluorid	Osteoporose	Erhöhung der Knochendichte [112]	Deutlicher Anstieg der Frakturrate [112]
Antiarrhythmika	Z. n. Myokardinfarkt	Reduktion von EKG-Auffälligkeiten [113]	Anstieg der Mortalität [114]
Interferon β	Multiple Sklerose	70%ige Reduktion neuer Hirnläsionen im MRT [115–118]	Keine überzeugende Wirkung hinsichtlich Behinderung [115–118]
Milrinon und Epoprostanol	Herzinsuffizienz	Verbesserung der Belastungstoleranz [119, 120]	Anstieg der Mortalität [121, 122]
Ibopamin	Herzinsuffizienz	Verbesserung der Ejektionsfraktion und Herzfrequenzvariabilität [123]	Anstieg der Mortalität [124]

EKG = Elektrokardiogramm

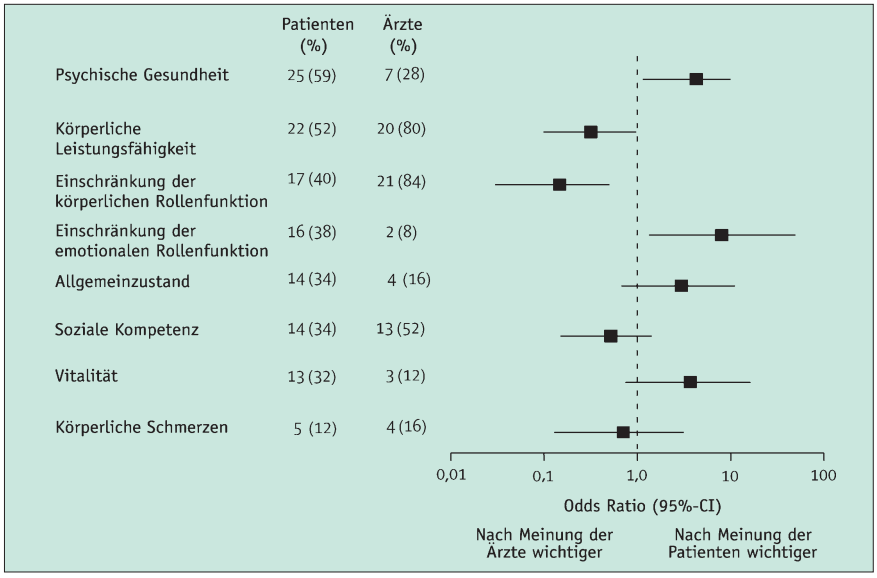

Abb. 5. Vergleich der wichtigsten Aspekte der Lebensqualität bei Multipler Sklerose nach Einschätzung von Patienten und Ärzten (auf der Basis von drei von acht Aspekten der gesundheitsbezogenen Lebensqualität, die anhand des Gesundheitsfragebogens SF-36 bewertet wurde).

nation von Endpunkten kann ein nützliches Maß für den Gesamteffekt einer Therapie auf alle relevanten Zielgrößen darstellen und sorgt meist für eine größere statistische Power, ist aber auch mit Problemen behaftet. So kann das für einen individuellen Patienten wichtigste Ergebnis durch die Therapie ganz anders beeinflusst sein als der kombinierte Endpunkt. Auch wenn der Thrombozytenaggregationshemmer Dipyridamol das Risiko des kombinierten Endpunktes aus Schlaganfall, Myokardinfarkt oder vaskulär bedingtem Tod senkt, scheint er sich nicht auf das alleinige Myokardinfarktrisiko auszuwirken [133] und wäre für einen Patienten mit instabiler koronarer Herzkrankheit nicht der optimale Wirkstoff. In kombinierten Endpunkten werden manchmal auch Ereignisse ganz unterschiedlicher Schweregrade zusammengefasst, und Therapieeffekte ergeben sich für den unwichtigsten Endpunkt, weil dieser oft am häufigsten vorkommt. Dies kommt zuweilen z. B. in Studien über Schlaganfallprophylaxe vor, bei denen in einem kombinierten Endpunkt auch transiente ischämische Attacken berücksichtigt werden. Ein ebenso problematischer kombinierter Endpunkt ist eine Mischung aus bestimmten klinischen Ereignissen und Krankenhausaufenthalten. Vermutlich wird die Tatsache, dass ein Patient an einem RCT teilnimmt, die Wahrscheinlichkeit einer Hospitalisierung beeinflussen, und mit Sicherheit wird diese Wahrscheinlichkeit je nach Gesundheitssystem auch Schwankungen unterworfen sein.

Behandlungsdauer und Nachbeobachtung

Eine weitere häufig vorkommende Schwierigkeit für die externe Validität von RCTs ergibt sich aus der unzureichenden Dauer der Behandlung und/oder Nachbeobachtung. Auch wenn z. B. Patienten mit refraktärer Epilepsie der jahrelangen Behandlung bedürfen, werden in den meisten RCTs über neue Medikamente die Wirkungen einer solchen Therapie nur über einige wenige Wochen beobachtet [131, 132]. Ob das initiale Ansprechen ein guter Prädiktor für einen Langzeitnutzen darstellt, ist nicht bekannt. Auf dasselbe Problem trifft man in RCTs über Schizophrenie: In weniger als 50% der Studien dauert die Nachbeobachtung länger als 6 Wochen, und in nur 20% werden die Patienten mehr als 6 Monate nachbeobachtet [33, 128]. Auf die Diskrepanz zwischen den nützlichen Behandlungswirkungen in Kurzzeit-RCTs und den weniger ermutigenden Erfahrungen von Langzeittherapien im klinischen Alltag wurde auch von Ärzten hingewiesen, die Patienten mit rheumatoider Arthritis behandeln [134].

Unerwünschte Behandlungswirkungen

Die Darstellung unerwünschter Behandlungswirkungen ist in RCTs und systematischen Reviews häufig ausgesprochen mangelhaft. In einem Review von 192 pharmazeutischen Studien ergab sich, dass die Darlegung unerwünschter klinischer Ereignisse oder toxischer Laborbefunde in weniger als einem Drittel der Studien angemessen war [135]. Die Therapieabbruchraten geben Hinweise auf die Verträglichkeit eines Medikaments, doch verwenden pharmazeutische Studien oftmals Einschlusskriterien und Run-in-Phasen, um Patienten auszuschließen, die für unerwünschte Wirkungen anfällig sein könnten. Im Praxisalltag sind daher oftmals höhere Therapieabbruchraten zu beobachten [136, 137]. Publikationsbias und die unzulängliche Darlegung unerwünschter Ereignisse in von der pharmazeutischen Industrie unterstützten RCTs stellen ein seit langem bestehendes und nach wie vor ungelöstes Problem dar [138, 139].

Die größten Sorgen bereitet Ärzten meist die externe Validität von RCTs über potenziell gefährliche Therapien. In den Industrienationen gehören iatrogene Komplikationen zu den führenden Mortalitätsursachen [140]. Vor allem bei der Einführung neuer Behandlungen, wenn für Studien oftmals schwer erkrankte Patienten rekrutiert werden, kann es zwar zur Überschätzung der Risiken kommen, doch ergeben sich aufgrund einer strengen Patientenselektion, der Begrenzung auf Spezialzentren und der intensiven Sicherheitsüberwachung meist geringere Risikoraten als im Klinikalltag. Ein gutes Beispiel sind RCTs über die Gabe von Warfarin bei nicht-rheumatischem Vorhofflimmern. Alle Studien berichteten über den Nutzen von Warfarin, doch im Rahmen der Studien traten deutlich geringere Komplikationsraten auf als im Praxisalltag [24, 141]. Daraus resultierende Zweifel an der externen Validität

sind zum Teil für die relevante Unterverschreibung von Warfarin verantwortlich zu machen, vor allem bei Patienten über 75 Jahren [24, 142, 143], die mehr als 70% der Fälle nicht-rheumatischen Vorhofflimmerns stellen [24] und unbehandelt das höchste Risiko aufweisen [24, 143].

Validität von routinemäßig erhobenen Daten

Oftmals wird davon ausgegangen, dass nichtrandomisierte Therapievergleiche auf der Grundlage routinemäßig erhobener Daten eine höhere externe Validität aufweisen als RCTs, da sie alle Patienten berücksichtigen, dem realen Praxisalltag entstammen und auch die Auswirkungen des Arzt-Patient-Verhältnisses und die Patientenpräferenzen berücksichtigen [144, 145]. Obschon bei angemessener Adjustierung nach Fallmix die Ergebnisse solcher Therapievergleiche denen von RCTs manchmal ähnlich sind [146, 147], kann man unmöglich sicher sein, dass systematische Verzerrungen (Bias) evident werden oder korrigierbar sind. Die in Tabelle 1 präsentierte Analyse von Patienten, die in der ECST-Studie für die Endarteriektomie randomisiert worden waren, hat gezeigt, dass sich hinter anscheinend ähnlichen klinischen Charakteristika größere Prognoseunterschiede verbergen können. Ferner wird dabei auch ein interessanter nichtrandomisierter Vergleich möglich, nämlich ein Vergleich des Outcomes innerhalb der für die Operation randomisierten Gruppe zwischen den operierten (Behandlungsgruppe) und den nicht-operierten Patienten (Kontrollgruppe). Der nichtrandomisierte Vergleich lässt darauf schließen, dass die Operation die Odds für Schlaganfall oder Tod nach fünf Jahren um mehr als die Hälfte senkt (Odds Ratio 0,46; 95%-CI 0,28 bis 0,77; $p = 0,003$; Abb. 6) und dieser Benefit bei Adjustierung nach Fallmix sogar noch größer ausfällt (0,32; 95%-CI 0,15 bis 0,57; $p < 0,001$).

Die entsprechende auf der randomisierten Behandlungszuteilung basierende Intention-to-treat-Analyse in der ECST-Studie ergibt jedoch, dass sich der Gesamtnutzen der Endarteriektomie tatsächlich aber nicht auf alle Stenosegrade erstreckt (1,02; 95%-CI 0,88 bis 1,18; $p = 0,90$; Abb. 6). Der nichtrandomisierte Vergleich ist natürlich konstruiert und unangebracht, da es, wie oben erläutert, spezifische Gründe dafür gab, warum Patienten, die für die Operation randomisiert worden waren, nicht operiert wurden; er veranschaulicht aber die Tatsache, dass eine solche Verzerrung durch eine Fallmix-Adjustierung nicht zuverlässig korrigiert werden kann [148]. Routinemäßig erhobene Daten sind dann von Nutzen, wenn die Durchführung von RCTs, etwa zur Beurteilung seltener unerwünschter Ereignisse, nicht praktikabel ist. Letztlich stellen sie aber eher ein Zusatzinstrument dar als eine Alternative.

Fazit und Empfehlungen

Randomisierte Studien und systematische Reviews liefern die zuverlässigsten Daten über Therapieeffekte, und zahlreiche schwerwiegende Irrtümer kamen zustande, weil man sich auf andere Arten von Evidenz stützte (Kasten 3). Auch wenn die dogmatische Weigerung von Ärzten, die Ergebnisse von RCTs anzuerkennen, nicht akzeptabel ist, so sind die Bedenken hinsichtlich der oftmals schwachen externen Validität doch durchaus berechtigt. Eine solche Einstellung führt im Praxisalltag zur Unterverschreibung von Therapien, die sich in Studien als wirksam erwiesen haben. Obschon manche Studien eine sehr gute externe Validität aufweisen [93, 149], trifft diese Feststellung, wie oben ausgeführt, auf viele – insbesondere auf die von der pharmazeutischen Industrie durchgeführten – Studien leider nicht zu. Und doch wird die angemessene Berücksichtigung der externen Validität von Forschern, Studienträgern, Ethikkommissionen, Fachzeitschriften, der pharmazeutischen Industrie und den Zulassungsbehörden gleichermaßen missachtet (Kasten 1). Das Urteil darüber bleibt den Ärzten überlassen, doch ist die Darlegung von Determinanten der externen Validität in Studienpublikationen, und hier vor allem in Sekundärveröffentlichungen und klinischen Leitlinien, nur selten ange-

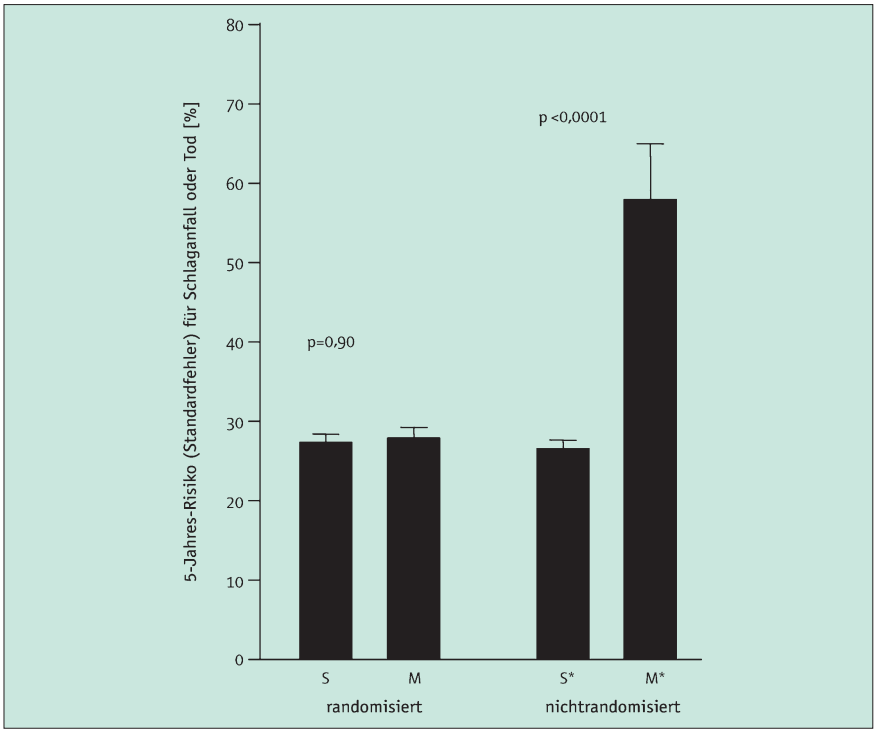

Abb. 6. 5-Jahres-Risiken für Schlaganfall oder Tod für alle Patienten, die im Rahmen der ECST-Studie [60] für die Operation (S) versus alleinige medikamentöse Therapie (M) randomisiert worden waren, und entsprechender nichtrandomisierter Vergleich innerhalb der für die OP randomisierten Patientengruppe, d.h. zwischen den Patienten, die operiert (S*) und denen, die nicht operiert wurden (M*). Ausgangscharakteristika siehe Tabelle 1.

messen. Manchmal sind Informationen in einem vorab veröffentlichten Methodenpapier zu finden, das nicht selten in irgendeiner obskuren Zeitschrift begraben und dem viel beschäftigten Arzt nicht leicht zugänglich ist, und ein Großteil der relevanten Informationen gelangt niemals zur Veröffentlichung. Man kann nicht davon ausgehen, dass RCTs und systematische Reviews Ergebnisse erbringen, die für alle Patienten und alle Settings unmittelbar relevant sind. Doch um auch extern valide zu sein, sollten sie zumindest so angelegt und dargestellt werden, dass Patienten und Ärzte sich ein Urteil darüber bilden können, auf wen sich die Ergebnisse sinnvoll anwenden lassen. Eingedenk der Gefahren einer Überregulation mögen einige der folgenden Empfehlungen bedenkenswert sein:

- Durchführung weiterer Untersuchungen zur externen Validität von RCTs, insbesondere in Verbindung mit dem untersuchten Therapieeffekt
- strengere Anforderungen als bisher an die externe Validität von RCTs, die bei den Zulassungsbehörden eingereicht werden
- verstärkte Berücksichtigung der externen Validität in den CONSORT-Leitlinien für die Abfassung von RCT-Studienberichten [49] und in den Leitlinien der Cochrane Collaboration für systematische Reviews [50] sowie Vereinbarung einer Checkliste, wie z. B. in Kasten 2 vorgeschlagen.
- Das Internationale Komitee der Herausgeber medizinischer Fachzeitschriften (International Committee of Medical Journal Editors) sollte verlangen, dass alle Primärberichte über RCTs oder systematische Reviews einen Abschnitt enthalten, der die Überschrift trägt: „Auf wen lassen sich diese Ergebnisse anwenden?".

Kasten 3: Beispiele für Interventionen, die als nützlich (bzw. schädlich) galten, sich in nachfolgenden RCTs aber als schädlich (bzw. nützlich) herausstellten

Für nützlich gehalten, aber nachweislich schädlich

- Hochdosierte Sauerstofftherapie bei Neugeborenen
- Antiarrhythmika nach Myokardinfarkt
- Fluoridbehandlung bei Osteoporose
- Bettruhe bei Zwillingsschwangerschaft
- Hormonersatztherapie zur Prävention von Gefäßerkrankungen
- Extra-/intrakranielle Bypassoperation zur Schlaganfallprophylaxe
- Hochdosierte Aspirintherapie bei Karotisendarteriektomie

Für schädlich gehalten, aber nachweislich nützlich

- β-Blocker bei Herzinsuffizienz
- Digoxin nach Myokardinfarkt

Literatur

[1] Cochrane AL. Effectiveness and Efficiency: random reflections on Health Services. London: Nuffield Provincial Hospitals Trust, 1972.

[2] Horton R. Common sense and figures: the rhetoric of validity in medicine (Bradford Hill Memorial Lecture 1999). Stat Med 2000;19:3149–64.

[3] Pocock SJ. Clinical trials: a practical approach. Chichester: John Wiley, 1983.

[4] Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 3rd edn. New York: Springer, 1998.

[5] Black D. The limitations of evidence. J R Coll Physicians Lond 1998;32:23–6.

[6] Hampton JR. Size isn't everything. Stat Med 2002;21:2807–14.

[7] Caplan LR. Evidence based medicine: concerns of a clinical neurologist. J Neurol Neurosurg Psychiatry 2001;71: 569–74.

[8] Evans JG. Evidence-based and evidence-biased medicine. Age Ageing 2 1995;4: 461–3.

[9] Charlton BG, Miles A. The rise and fall of EBM. Q J Med 1998;91:371–4.

[10] Naylor C. Grey zones in clinical practice: some limits to evidence-based medicine. Lancet 1995;345:840–2.

[11] Swales JD. Evidence-based medicine and hypertension. J Hypertens 1999; 17:1511–6.

[12] Morgan WKC. On evidence, embellishment and efficacy. J Eval Clin Pract 1997;3:117–22.

[13] Feinstein AR, Horwitz RI. Problems in the "evidence" of "evidence-based medicine" Am J Med 1997;103: 529–35.

[14] Pashos CL, Normand ST, Garfinkle JB, Newhouse JP, Epstein AM, McNeil B. Trends in use of drug therapies in patients with acute myocardial infarction: 1988–1992. J Am Coll Cardiol 1994; 23:1023–30.

[15] Garfield FB, Garfield JM. Clinical judgement and clinical practice guidelines. Int J Technol Assess Health Care 2000; 16:1050–60.

[16] Cabana MB, Rand CS, Powe NR, et al. Why don't clinicians follow clinical practice guidelines? A framework for improvement. JAMA 1999;282:1458–65.

[17] Davis DA, Taylor-Vaisey A. Translating guidelines into practice, a systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines. Can Med Assoc J 1997;157: 408–16.

[18] Ford LG, Hunter CP, Diehr P, Frelick RW, Yates J. Effect of patient management guidelines on physician practice patterns: the community hospital oncology program experience. J Clin Oncol 1987; 5:504–11.

[19] Grol R, Dalhuijsen J, Thomas S, et al. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. BMJ 1998; 317:858–61.

[20] Messerli FH. Antihypertensive therapy: β-blockers and diuretics – why do physicians not always follow guidelines? Proceedings/Baylor University Medical Center 2000;13:128–31.

[21] Sonis J, Doukas D, Klinkman M, Reed B, Ruffin MT. Applicability of clinical trial results to primary care. J Am Med Soc 1998;280:1746.

[22] Wilson S, Delaney BC, Roalfe A, et al. Randomized controlled trials in primary care: case study. BMJ 2000;321:24–7.

[23] Sellors J, Cosby R, Trim K, et al, for the Seniors Medication Assessment Research Trial (SMART) Group.Recruiting family physicians and patients for a clinical trial: lessons learned. Fam Pract 2002;19:99–104.

[24] Oswald N, Bateman H. Applying research evidence to individuals in primary care: a study using non-rheumatic atrial fibrillation. Fam Pract 1999;16: 414–9.

[25] Fahey T. Applying the results of clinical trials to patients in general practice: perceived problems, strengths, assumptions, and challenges for the future. Br J Gen Pract 1998;48: 1173–8.

[26] Mant D. Can randomised trials inform clinical decisions about individual patients? Lancet 1999;353:743–6.

[27] Jacobson LD, Edwards AGK, Granier SK, Butler CC. Evidence-based medicine and general practice. Br J Gen Pract 1997;47:449–52.

[28] McCormick JS. The place of judgement in medicine. Br J Gen Pract 1994;44: 50–1.

[29] Benech I, Wilson AE, Dowell AC. Evidence-based practice in primary care: past, present and future. J Eval Clin Prac 1996;2:249–63.

[30] Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Threats to applicability of randomised trials: ex-

clusions and selective participation. J Health Serv Res Policy 1999;4: 112–21.
[31] Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998;52: 377–84.
[32] Egglin TKP, Horwitz RI. The case for better research standards in peripheral thrombolysis: poor quality of randomised trials during the past decade. Acad Radiol 1996;3:1–9.
[33] Gilbody S, Wahlbeck K, Adams C. Randomized controlled trials in schizophrenia: a critical perspective on the literature. Acta Psychiatr Scand 2002;105: 243–51.
[34] Licht RW, Gouliaev G, Vestergaard, Frydenberg M. Generalisability of results of randomised drug trials. Br J Psychiatry 1997;170:264–7.
[35] Camm AJ. Clinical trials of arrhythmia management: methods or madness. Control Clin Trials 1996;17: 4s–16s.
[36] Norris SL, Engelgau MM, Narayan KMV. Effectiveness of self-management training in type 2 diabetes: a systematic review of randomised controlled trials. Diabetes Care 2001;24: 561–87.
[37] Moore DAJ, Goodall RL, Ives NJ, Hooker M, Gazzard BG, Easterbrook PJ. How generalisable are the results of large randomised controlled trials of antiretroviral therapy. HIV Med 2000;1: 149–54.
[38] Brown N, Melville M, Gray D, 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 1999;81:598–602.
[39] Charleson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. BMJ 1984;289:1281–4.
[40] Simon GE, Vonkorff M, Heilingenstein JH, et al. Initial antidepressant choice in primary care: effectiveness of fluoxetine vs. tricyclic antidepressants. JAMA 1996;275:1897–902.
[41] Davey Smith G, Egger M. Incommunicable knowledge? Interpreting and applying the results of clinical trials and meta-analyses. J Clin Epidemiol 1998; 51: 289–95.
[42] Hennen BK. Measuring complexity of clinical problems. J Med Educ 1984;59: 487–93.
[43] Julian DG, Pocock SJ. Interpreting a trials report. In Pitt B, Julian D, Pocock S, eds. Clinical Trials in Cardiology London: WB Saunders, 1997: 33–42.
[44] Idanpaan-Heikkila JE. WHO guidelines for good clinical practice (GCP) for trials on pharmaceutical products: responsibilities of the investigator. Ann Med 1994;26: 89–94.
[45] Wermeling DP. Clinical research: regulatory issues. Am J Health Syst Pharm 1999;56: 252–6.
[46] Medical Research Council. MRC Guidelines for Good Clinical Practice in Clinical Trials. London: Medical Research Council , 1998.
[47] Medical Research Council. Clinical Trials for Tomorrow. London: Medical Research Council, 2003.
[48] Governance arrangements for NHS Research Ethics Committees. www.dh.gov.uk/assetRoot/04/05/86/09/04058609.pdf (assessed July 21, 2004)
[49] Altman DG, Schulz KF, Moher D, et al, for the Consort Group. The revised CONSORT statement for reporting randomised trials: explanation and elaboration. Ann Intern Med 2001,134: 663–94.
[50] Alderson P, Green S, Higgins JPT, eds. Cochrane Reviewers' Handbook 4.2.1] [updated December 2003]. In: The Cochrane Library, Issue 1, 2004. Chichester: John Wiley & Sons Ltd.
[51] LeBaron S, Reyher J, Stack JM. Paternalistic vs egalitarian physician styles: the treatment of patients in crisis. J Fam Pract 1985;21:56–6.2
[52] Thomas KB. General practice consultations: is there any point in being positive? BMJ 1987;294:1200–2.
[53] Di Blasi Z, Harkness E, Ernst E, Georgiou A, Kleijnen J. Influence of context effects on health outcomes: a systematic review. Lancet 2001;357:757–62.
[54] Kleijnen J, de Craen AJM, van Everdingen J, Krol L. Placebo effect in double-blind clinical trials: a review of interactions with medications. Lancet 1994; 344: 1347–9.
[55] Kaptchuk TJ. Powerful placebo: the dark side of the randomised controlled trial. Lancet 1998;351:1722–5.
[56] Onel E, Hammond C, Wasson JH, et al. An assessment of the feasibility and impact of shared decision making in prostate cancer. Urology 1998;51: 63–6.
[57] Benson J, Britten N. Patients decisions about whether or not to take antihypertensive drugs: qualitative study. BMJ 2002;325:873–6.
[58] Redelmeier DA, Rozin P, Kahneman D. Understanding patients' decisions. JAMA 1993;270:72–6.
[59] Olschewski M, Schumacher M, Davis KB. Analysis of randomised and non-randomized patients in clinical trials using the comprehensive cohort follow-up study design. Control Clin Trials 1992;13:226–39.
[60] European Carotid Surgery Trialists' Collaborative Group. Randomised trial of endarterectomy for recently symptomatic carotid stenosis: final results of the MRC European Carotid Surgery Trial (ECST). Lancet 1998;351:1379–87.
[61] Lovett JK, Coull A, Rothwell PM, on behalf of the Oxford Vascular Study. Early risk of recurrent stroke by aetiological subtype: implications for stroke prevention. Neurology 2004;62:569–74.
[62] Masuhr F, Busch M, Einhaupl KM. Differences in medical and surgical therapy for stroke prevention between leading experts in North America and Western Europe. Stroke 1998;29: 339–45.
[63] Sacco RL, Kargman DE, Gu Q, Zamanillo MC. Race-ethnicity and determinants of intracranial atherosclerotic cerebral infarction. The Northern Manhattan Stroke Study. Stroke 1995;26: 14–20.
[64] Fine PEM. Variation in protection by BCG: implications of and for heterologous immunity. Lancet 1995;346: 1339–45.
[65] Ricci S, Celani MG, Righetti E, Cantisani AT, for the International Stroke Trial Collaborative Group. Between country variations in the use of medical treatments for acute stroke: An update. Cerebrovasc Dis 1996;6 (Suppl 2):133.
[66] Roberts C. The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. Stat Med 1999;18: 605–15.
[67] Downs SH, Black NA, Devlin HB, Royston CMS, Russell RCG. Systematic review of the effectiveness and safety of laparoscopic cholecystectomy. Ann R Coll Surg Engl 1996;78:241–323.
[68] Black NA, Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. Br J Urol 71996;8:497–510.
[69] Asymptomatic Carotid Atherosclerosis Study Group. Carotid endarterectomy for patients with asymptomatic internal carotid artery stenosis. JAMA 1995;273:1421–8.
[70] Moore WS, Young B, Baker WH, et al. Surgical results: a justification of the surgeon selection process for the ACAS trial. J Vasc Surg 1996;23:323–38.
[71] Bond R, Rerkasem K, Rothwell PM. High morbidity due to endarterectomy for asymptomatic carotid stenosis. Cerebrovasc Dis 2003;16 (Suppl 4):65.
[72] Barnett HJM, Barnes RW, Clagett GP, Ferguson GG, Robertson JT, Walker PM. Symptomatic carotid artery stenosis: a

solvable problem. The NASCET trial. Stroke 1992;23:1048–53.
[73] De Vita VT. Breast cancer therapy; exercising all our options. N Engl J Med 1989;320:527–9.
[74] Gurwitz JH, Col NF, Avorn J. The exclusion of elderly and women from clinical trials in acute myocardial infarction. JAMA 1992;268:1417–22.
[75] Bungeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ 1997;315:1059.
[76] Jones R, Lydeard S. Prevalence of symptoms of dyspepsia in the community. BMJ 1989;298:30–2.
[77] Kay L. Prevalence, incidence and prognosis of gastrointestinal symptoms in a random sample of an elderly population. Age Ageing 1994;23:146–9.
[78] Jorgensen HS, Nakayama H, Kammersgaard LP, et al. Predicted impact of intravenous thrombolysis on prognosis of general population of stroke patients: simulation model. BMJ 1999;319: 288–9.
[79] National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischaemic stroke. N Engl J Med 1995;333:1581–7.
[80] LaRue LJ, Alter M, Traven ND, et al. Acute stroke therapy trials: problems in patient accrual. Stroke 1998;19: 950–4.
[81] Maynard C, Althouse R, Cerqueira M, Olsufka M, Kennedy JW. Underutilisation of thrombolytic therapy in eligible women with acute myocardial infarction. Am J Cardiol 1991;68:529–30.
[82] Steiner TI, Clifford Rose F. Towards a model stroke trial. The single centre naftidrofuryl study. Neuroepidemiology 1986;5:121–47.
[83] Rovers MM, Zielhuis GA, Bennett K, Haggard M. Generalisability of clinical trials in otitis media with effusion. Int J Pediatr Otorhinolaryngol 2001;60: 29–40.
[84] Califf RM, Pryor DB, Greenfield JC. Beyond randomised clinical trials: applying clinical experience in the treatment of patients with coronary artery disease. Circulation 1986;74: 1191–4.
[85] Stroke Prevention in Atrial Fibrillation Investigators. Stroke Prevention in Atrial Fibrillation Study. Circulation 1991; 84:527–39.
[86] Muller DWM. Selection of patients with acute myocardial infarction for thrombolytic therapy. Ann Intern Med 1990; 113:949–60.
[87] Henderson RA, Raskino CL, Hampton R. Variations in the use of coronary arteriography in the UK: the RITA trial coronary arteriogram register. Q J Med 1995;88:167–73.
[88] Bjom M, Brendstrup C, Karlsen S, Carlsen JE. Consecutive screening and enrolment in clinical trials: the way to representative patient samples? J Card Fail 1998;4:225–30.
[89] Bowen JT, Barnes TRE. The clinical characteristics of schizophrenic patients consenting or not consenting to a placebo controlled trial of antipsychotic medication. Hum Psychopharmacol 1994;9:432–3.
[90] Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. Stat Med 1996;15:263–71.
[91] Stiller CA. Centralised treatment, entry to trials and survival. Br J Cancer 1994; 70:352–62.
[92] Woods Kl, Ketley D. lntravenous β blockade in acute myocardial infarction. Doubt exists about external validity of trials of intravenous β blockade. BMJ 1999;318:328–9.
[93] Gruppo Italiano per 10 Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Lancet 1986;1: 397–402.
[94] Pablos-Mendez A, Barr RG, Shea S. Run-in periods in randomised trials: implications for the application of results in clinical practice. JAMA 1998; 279:222–5.
[95] Haynes RB, Dantes R. Patient compliance and the conduct and interpretation of therapeutic trials. Control Clin Trials 1987;8:12–9.
[96] Gomez-Marin O, Prineas RJ, Sinaiko AR. The sodium-potassium blood pressure trial in children: design, recruitment and randomisation. Control Clin Trials 1991;12:408–23.
[97] Australia-New Zealand Heart Failure Research Collaborative Group. Effects of carvedilol, a vasodilatory β-blocker, in patients with congestive heart failure due to ischaemic heart disease. Circulation 1995;92:212–8.
[98] Packer M, Bristow MR, Cohn JN et al, for the US Carvedilol Heart Failure Study Group. The effects of carvedilol on morbidity and mortality in patients with chronic heart failure. N Engl J Med 1996;334:1349–55.
[99] Amery W, Dony J. A clinical trial design avoiding undue placebo treatment. J Clin Pharmacol 1975;15:674–9.
[100] Quitkin FM, Rabkin JG. Methodological problems in studies of depressive disorder: utility of the discontinuation design. J Clin Psychopharmacol 1981; 1:283–8.
[101] Hallstrom A, Verter J, Friedman L. Randomising responders. Control Clin Trials 1991;12:486–503.
[102] Leber PD, Davis CS. Threats to the validity of clinical trials employing enrichment strategies for sample selection. Control Clin Trials 1998;19: 178–87.
[103] Davis KL, Thai LJ, Gamzu ER, et al, and the Tacrine Collaborative Study Group. A double-blind, placebo-controlled multicenter study of tacrine for Alzheimer's disease. N Engl J Med 1992;327:1253–9.
[104] Hall JC, Mills B, Nguyen H, Hall JL. Methodologic standards in surgical trials. Surgery 1996;119:466–72.
[105] Shapiro SH, Weijer C, Freedman B. Reporting the study populations of clinical trials. CIear transmission or static on the line? J Clin Epidemiol 2000;53: 973–9.
[106] European Atrial Fibrillation Trial (EAFT) Study Group. Secondary prevention in non-rheumatic atrial fibrillation after transient ischaemic attack or minor stroke. Lancet 1993;342: 1255–62.
[107] Algra A, Francke CL, Koehler PJJ, for the Stroke Prevention in Reversible Ischaemia Trial (SPIRIT) group. A randomized trial of anticoagulants versus aspirin after cerebral ischaemia of presumed arterial origin. Ann Neurol 1997;42:857–65.
[108] Gorter JW, for the Stroke Prevention in Reversible Ischaemia Trial (SPIRIT) and European Atrial Fibrillation Trial (EAFT) groups. Major bleeding during anticoagulation after cerebral ischaemia: patterns and risk factors. Neurology 1999;53:1319–27.
[109] Rothwell PM, Eliasziw M, Gutnikov SA, et al, for the Carotid Endarterectomy Trialists' Collaboration. Pooled analysis of individual patient data from randomised controlled trials of endarterectomy for symptomatic carotid stenosis. Lancet 2003;361: 107–16.
[110] Merlo J, Broms K, Undblad U, et al. Association of outpatient utilisation of non-steroidal anti-inflammatory drugs and hospitalised heart failure in the entire Swedish population. Eur J Clin Pharmacol 2001;57:71–5.
[111] Rothwell PM, Slattery J, Warlow CP. A systematic review of the risks of stroke and death due to carotid endarterectomy for symptomatic stenosis. Stroke 1996;27:260–5.
[112] Riggs BL, Hodgson SF, O'Fallon WM, et al. Effect of fluoride treatment on fracture rate in postmenopausal women with osteoporosis. N Engl J Med 1990;322:802–9.
[113] McAlistair FA, Teo KK. Antiarrhythmic therapies for the prevention of sudden cardiac death. Drugs 1997;54: 235–52.

[114] The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 19898;321:406–12.
[115] The IFNB Multiple Sclerosis Study Group. Interferon β-lb is effective in relapsing-remitting multiple sclerosis. 1. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. Neurology 1993;43: 655–61.
[116] The IFNB Multiple Sclerosis Study Group and the University of British Columbia MS/MRI Analysis Group. Interferon β-lb in the treatment of multiple sclerosis: final outcome of the randomized controlled trial. Neurology 1995; 45: 1277–85
[117] Jacobs LD, Cookfair DL, Rudick AR, et al. Intramuscular interferon β-la for disease progression in relapsing multiple sclerosis. Ann Neurol 1996;39: 285–94.
[118] Ebers GC for the PRISMS Study Group. Randomised double-blind placebo-controlled study of interferon β-la in relapsing-remitting multiple sclerosis. Lancet 1998;352: 1498–504.
[119] Di Bianco R, Shabetai R, Kostuk W, et al. A comparison of oral milrinone, digoxin and their combination in the treatment of patients with chronic heart failure. N Engl J Med 1989;320: 677–83.
[120] Sueta CA, Gheorghiade M, Adams KF, et al, and the Epoprostenol Multicentre Research Group. Safety and efficacy of epoprostenol in patients with severe congestive heart failure. Am J Cardiol 1995;75:34A–43A.
[121] Packer M, Carver JR, Rodeheffer RJ, et al, for the Promise Study Research Group. Effect of oral milrinone on mortality in severe chronic heart failure. N Engl J Med 1991;325:1468–75.
[122] Califf RM, Adams KF, McKenna WJ, et al. A randomized controlled trial of epoprostenol therapy for severe congestive heart failure. Am Heart J 1997;134:44–54.
[123] Yee KM, Struthers AD. Can drug effects on mortality in heart failure be predicted by any surrogate outcome measure? Eur Heart J 1997;18:1860–4.
[124] Hampton JR, van Veldhuisen DJ, Kleber FX, et al (1997) Randomised study of ibopamine on survival in patients with advanced severe heart failure. Lancet 349: 971–77
[125] Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. Stat Med 1995;14:331–45.
[126] van Gijn J, Warlow C. Down with stroke scales. Cerebrovasc Dis 1992; 2:244–6.
[127] Goetzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of non-steroidal antiinflammatory drugs in rheumatoid arthritis. Control Clin Trials 1989;10:31–56.
[128] Thornley B, Adams CE. Content and quality of 2000 controlled trials in schizophrenia over 50 years. BMJ 1998;317:1181–4.
[129] Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales – a major source of bias in randomised controlled trials of treatments for schizophrenia? Br J Psychiatry 2000;176:249–52.
[130] Rothwell PM, McDowell Z, Wong CK, P Dorman. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ 1997;314:1580–3.
[131] Binnie CD. Design of clinical antiepileptic drug trials. Seizure 1995; 4:187–92.
[132] Walker MC, Sander JW. Difficulties in extrapolating from clinical trial data to clinical practice: the case of antiepileptic drugs. Neurology 1997; 49:333–7.
[133] Diener HC, Cunha L, Forbes C, Sivenius J, Smets P, Lowenthal A, for the European Stroke Prevention Study. 2. Dipyridamole and acetylsalicylic acid in the secondary prevention of stroke. J Neurol Sci 1996;143:1–13.
[134] Pincus T. Rheumatoid arthritis: disappointing long-term outcomes despite successful short-term clinical trials. J Clin Epidemiol 1998;41:1037–41.
[135] Ioannidis JP, Contopoulos-Ioannidis DG. Reporting of safety data from randomised trials. Lancet 1998;352: 1752–3.
[136] Jones J, Gorkin L, Lian J, Staffa JA, Fletcher AP. Discontinuation of and changes in treatment after start of new courses of antihypertensive drugs: a study of a United Kingdom population. BMJ 1995;311:293–5.
[137] Andrade SE, Walker AM, Gottlieb LK, et al. Discontinuation of antihyperlipidaemic drugs – do rates reported in clinical trials reflect rates in primary care settings. N Engl J Med 1995; 332:1125–31.
[138] Hemminki E. Study of information submitted by drug companies to licensing authorities. BMJ 1980;280: 833–6.
[139] McPherson K, Hemminki E. Synthesising licensing data to assess drug safety. BMJ 2004;328:518–20.
[140] Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalised patients. JAMA 1998; 279:1200–5.
[141] Landefeld CS, Beyth RJ. Anticoagulant-related bleeding: clinical epidemiology, prediction, and prevention. Am J Med 1993;95:315–28.
[142] Up GH, Golding DJ, Masood N, Beevers DG, Child DL, Fletcher RI. A survey of arial fibrillation in general practice. Br J Gen Pract 1997;47:285–9.
[143] Antani MR, Beyth RJ, Covinsky KE, et al. Failure to prescribe warfarin to patients with non-rheumatic atrial fibrillation. J Gen Intern Med 1996;11: 713–20.
[144] Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312: 1215–8.
[145] McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Methods in health service research. Interpreting the evidence: choosing between randomised and non-randomised studies. BMJ 1999;319:312–5.
[146] Concato J, Shah N, Horwitz RI. Randomised controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000; 342:1887–92.
[147] Benson K, Hartz AJ. A comparison of observational studies and randomized controlled trials. N Engl J Med 2000; 342:1878–86.
[148] Coronary Drug Project Research Group. Influence of adherence to treatment and response to cholesterol on mortality in the Coronary Drug Project. N Engl J Med 1980;303: 1038–41.
[149] Heart Protection Study Collaborative Group. MRCf/BHF Heart Protecction Study of cholesterol lowering with simvastatin in 20.536 high-risk indivuals: a randomised placebo-controlled trial. Lancet 2002;360:7–22.

Korrespondenzadresse:
Peter M. Rothwell
peter.rothwell@dneuro.ox.ac.uk

Anmerkung der Redaktion:
Die Übersetzung dieses Artikels erfolgte durch Frau Karin Beifuss (Stuttgart), die fachliche Bearbeitung übernahm Frau Gerta Rücker (IMBI – Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Freiburg). Beiden sei an dieser Stelle sehr herzlich gedankt.