

RANDOMISIERTE KONTROLLIERTE STUDIEN

Kritische Evaluation ist ein Wesensmerkmal ärztlichen Handelns

Die gegen randomisierte kontrollierte Studien vorgebrachten Einwände sind nicht überzeugend und zeigen oft Informationsdefizite.

Jürgen Windeler, Gerd Antes, Johann Behrens,
Norbert Donner-Banzhoff, Monika Lelgemann

Die drei Buchstaben RCT haben sich zu einem Reizwort in der Diskussion um die Nutzenbewertung entwickelt, auch im Deutschen Ärzteblatt (1–3). Das Kürzel RCT steht für „Randomized Controlled Trial“. Es beschreibt ein Studiendesign, in dem zwei (oder mehr) Gruppen von Patienten zeitgleich bezüglich der Ergebnisse zweier (oder mehr) Behandlungen verglichen werden sollen, wobei die Patienten diesen Behandlungen zufällig zugewiesen werden.

Zum Reizthema werden RCT natürlich nicht wegen des nüchternen methodischen Vorgehens. Der Grund dafür ist vielmehr, dass Studien mit Zufallszuteilung als die aussagekräftigsten für die Wirksamkeits- oder Nutzenbewertung angesehen werden. Vielfach wird die randomisierte Studie als Königsweg in der Bewertung der Vor- und Nachteile von medizinischen Interventionen bezeichnet. Bei Fehlen solcher Studien wird der Nutzen einer Behandlung kritisch bewertet. Folgerichtig ist es international üblich, solche Maßnahmen nicht (uneingeschränkt) zu empfehlen und diesbezügliche Aussagen, zum Beispiel in Leitlinien, zurückhaltend zu formulieren. Arzneimittel erhalten

ohne Wirksamkeitsnachweis aus solchen Studien keine Zulassung.

Es gibt fundierte theoretische Argumente und zahlreiche Beispiele dafür, dass RCT-Ergebnisse zu skeptischeren Bewertungen führen als die Ergebnisse anderer Studientypen oder der ärztlichen Erfahrung. Fehlerquellen wirken bevorzugt in Richtung fälschlich günstiger Ergebnisse. Aktuelle Beispiele dafür sind die Geschichte der Hormontherapie in der Postmenopause, der Vitaminsubstitution (4, 5, 6) und des Einsatzes von Rechtsherzkathetern (7). Es lohnt sich daher, einen nüchternen Blick auf randomisierte Studien, ihre Grundlagen und Grenzen zu werfen.

Randomisierung: entscheidende Vorteile

Im Kern geht es um die Frage, ob ein Patient mit der Anwendung eines Verfahrens besser „dran“ ist als ohne Anwendung dieses Verfahrens. Die Antwort ergibt sich aus dem Vergleich zwischen mindestens zwei Möglichkeiten. Das Ergebnis erlaubt eine kausale Aussage in dem Sinne, dass die Wahrscheinlichkeit für ein bestimmtes Ergebnis durch die Behandlung verändert wird.

Eine solche Aussage ist mit einem einzelnen Patienten nicht zuverlässig

und verallgemeinerbar zu treffen, unter anderem deshalb, weil eine gleichzeitige Anwendung zweier Alternativen offensichtlich unmöglich ist und eine Veränderung des Krankheitszustands und damit eine Änderung des Ausgangszustands für eine weitere Behandlung einen fairen Vergleich der Verfahren behindern.

Aussagen zum Nutzen eines medizinischen Behandlungsverfahrens stützen sich daher auf den Vergleich von zwei zeitgleich beobachteten Patientengruppen (mit und ohne Anwendung des Verfahrens). Diese beiden Patientengruppen dienen also als Ersatz für die Beobachtung eines einzelnen Patienten in zwei Situationen.

Führt man diesen Gedanken fort, so dürfen sich die Patientengruppen nicht systematisch unterscheiden. Generell sind für einen fairen Leistungsvergleich gleiche Ausgangs- und Rahmenbedingungen erforderlich (das Prinzip des „ceteris paribus“). Die Gleichheit der Ausgangsbedingungen, die man als Strukturgleichheit bezeichnet, bezieht sich auf alle Merkmale zu Beginn einer Studie.

Man kann versuchen, Strukturgleichheit durch spezielle Design-techniken „herzustellen“, etwa durch ein Matching von Studienteil-

Medizinischer Dienst der Spitzenverbände der Krankenkassen, Essen:
Prof. Dr. med. Windeler
Deutsches Cochrane-Zentrum, Freiburg:
Dr. rer. nat. Antes

Institut für Gesundheits- und Pflegewissenschaft, Martin-Luther-Universität Halle-Wittenberg:
Prof. Dr. phil. habil. Behrens

Abteilung für Allgemeinmedizin, Präventive und Rehabilitative Medizin, Philipps-Universität Marburg: Prof. Dr. med. Donner-Banzhoff MHS
Interdisziplinäres HTA-Zentrum in der Universität Bremen, c/o Institut für Gesundheits- und Medizinrecht: Lelgemann

nehmern oder andere aufwendigere Verfahren (8). Die Strukturgleichheit kann auch angestrebt werden durch die statistische Berücksichtigung von Unterschieden zwischen den Gruppen. Für diese sogenannte Adjustierung von Störfaktoren (Confounder) stehen verschiedene Standardverfahren zur Verfügung. Ein qualitativ hochwertiges Vorgehen ist zur „Herstellung“ der Strukturgleichheit in nicht randomisierten Studien zwar unabdingbar, jedoch in der Studienrealität selten anzutreffen; denn es ist:

- anspruchsvoll wegen der Notwendigkeit, das Vorgehen inklusive der einzubeziehenden Merkmale vorab in einem Studienprotokoll genau festzulegen und zu begründen
- aufwendig wegen der Notwendigkeit, zahlreiche Patientenmerkmale zu dokumentieren
- schwierig wegen der eingeschränkten Messbarkeit relevanter Merkmale (zum Beispiel Motivation).

Bei der Berücksichtigung nicht erhebbarer oder unbekannter Merkmale stößt jedes Verfahren, das versucht, solche Merkmale zu berücksichtigen, an prinzipielle Grenzen. Ein Studienergebnis erlaubt es dann nicht, den Behandlungseffekt von solchen störenden Merkmalen zuverlässig zu trennen. Ob das damit verbundene Risiko für Fehler in den Studienergebnissen in Kauf genommen werden kann, ist von Fall zu Fall zu entscheiden.

Die Randomisierung hat im Vergleich hierzu enorme Vorteile. Die Strukturgleichheit ist ein „Abfallprodukt“ der zufälligen Zuteilung. Es muss kein Merkmal bekannt sein, damit die Randomisierung gleiche Ausgangsbedingungen schafft (9).

Da die Kenntnis von Merkmalen nicht erforderlich ist, ist es auch unerheblich, ob es sich um bekannte oder unbekannte Merkmale handelt. Die Randomisierung sichert die Strukturgleichheit auch für nicht messbare und nicht bekannte Einflussfaktoren. Dies ist durch kein anderes Verfahren zu erreichen.

Die Erhebung zahlreicher Merkmale zur Beschreibung der Ausgangssituation erfordert einen beträchtlichen Aufwand, der bei einer Randomisierung drastisch reduziert

werden kann. Diesbezüglich sind randomisierte Studien also einfacher durchzuführen als Studien, in denen die Strukturgleichheit hergestellt werden muss. Betrachtet man diese Vorteile zusätzlich zu dem Umstand, dass Aussagen aus randomisierten Studien weniger fehleranfällig sind, so wird verständlich, warum die Randomisierung international als zentrales Prinzip für aussagefähige Studien anerkannt ist.

Bei diesen Vorteilen stellt sich die Frage, welche Argumente gegen die Durchführung einer prospektiv geplanten vergleichenden Studie mit Randomisierung vorgebracht werden (beispielsweise in 10). Hier sollen nicht die Sonder-situationen thematisiert werden, in denen solche Studien nicht durchgeführt werden müssen. Gründe können etwa sehr ausgeprägte Therapieeffekte sein oder auch quasi deterministische Verläufe wie die Durchführung einer Anästhesie. Es soll hier auch nicht um die Frage gehen, dass in besonderen Ausnahmesituationen zwar eine Randomisierung nicht durchführbar ist (wegen ausgeprägter Präferenzen der Patienten oder Behandler), aber ohne Weiteres eine prospektiv geplante vergleichende Studie.

Argumente gegen RCT, die man in der Diskussion häufig antrifft, lauten:

1. RCT sind nicht für alle Fragestellungen geeignet.

Diese Aussage ist ebenso richtig wie trivial. Für jede Fragestellung muss das geeignete Studiendesign gewählt werden. RCT sind für eine Frage, nämlich die nach der kausalen Beziehung zwischen einer Intervention und dem Ergebnis, der „Königsweg“. Auf diese Weise aber die Häufigkeit von Erkrankungen zu ermitteln, wäre ein Kunstfehler.

Der Eindruck, dass RCT (zu) häufig thematisiert werden, beruht nicht auf deren allumfassendem Anspruch, sondern dem großen Interesse an kausalen Zusammenhängen, entsprechenden Fragestellungen und Studien.

Die Eignung von RCT zur Beantwortung kausaler Fragestellungen ist unabhängig davon, wie diese

kausalen Beziehungen genannt werden. Die Unterscheidung in efficacy („Wirksamkeit unter Studienbedingungen“) und effectiveness („Wirksamkeit unter Alltagsbedingungen“) ist an sich schon nicht unproblematisch, da eine genaue Abgrenzung von Studien- und Alltagsbedingungen in der Praxis nicht möglich ist. Aus den beiden Begriffen unterschiedliche Forschungsmethoden abzuleiten, insbesondere für die Ermittlung der effectiveness andere Designs außer prospektiv vergleichenden Interventionsstudien zu favorisieren, ist abwegig. Bei efficacy und effectiveness handelt es sich um kausale Beziehungen zwischen Intervention und Ergebnis, was dann das prioritär zu wählende Studiendesign bestimmt.

2. Die Randomisierung führt nicht automatisch zu fehlerfreien Studien.

Auch diese Aussage ist richtig und trivial. In prospektiven vergleichenden Interventionsstudien ist die Randomisierung eines von mehreren methodischen Instrumenten, mit denen die Aussagekraft von Ergebnissen erhöht werden kann. Es gilt jedenfalls, dass eine Randomisierung „automatisch“ (das heißt im Mittel) zu einer Strukturgleichheit führt, was für nicht randomisierte Studien nicht gilt. Wird aber eine unangemessene oder falsch dosierte Therapie verwendet, dann sind Studien grundsätzlich nicht aussagefähig, unabhängig davon, ob sie randomisiert sind oder nicht (11, 12).

3. RCT sind bei seltenen Erkrankungen nicht durchführbar.

Diese Aussage gilt so pauschal zweifellos nicht, wie Beispiele zeigen (13, 14). Zum einen sind viele „seltene“ Erkrankungen nach der einschlägigen EU-Definition mit bis zu sechsstelligen Patientenzahlen in der EU häufig genug, um mehrere aussagefähige Studien durchzuführen. Zum anderen würde dieses Argument, wenn man es für RCT gelten ließe, auch für viele andere Studientypen, insbesondere auch für prospektive vergleichende Studien insgesamt gelten. Bei sehr seltenen Erkrankungen (Anhaltspunkt: weniger als 100 Fälle EU-weit)

können andere Wege, zum Beispiel aussagefähige Register oder N-of-1-Studien, beschritten werden, um zu Erkenntnissen zu kommen (15).

4. Randomisierte Studien sind (zu) teuer, (zu) aufwendig und nicht lange genug durchführbar.

Dies stimmt natürlich in absoluter Betrachtung (der Aufwand für eine Studie ist höher als der Aufwand ohne Studie). Es mag auch noch im Vergleich zu einzelnen Fallserien zutreffen. Im Vergleich zu aussagefähigen vergleichenden Studien ohne Randomisierung ist der Aufwand nicht höher, sondern oft sogar geringer. Die erforderliche Dauer der Nachbeobachtungszeit ist von der randomisierten Behandlungszuteilung unabhängig. Auch für Registerdaten muss zwischen der Anwendung einer Maßnahme und der Beobachtung eines (Ziel-)Ereignisses die für eine Bewertung erforderliche Zeit vergangen sein.

5. RCT sind ethisch nicht vertretbar.

Es muss daran erinnert werden, dass eine ganz entscheidende ethische Anforderung an Studien die ist, aussagefähige Ergebnisse zu liefern und damit dem persönlichen

sammen mit dem Patienten zu entscheiden, ob mit den aus RCT gelieferten validen Ergebnissen in der Behandlungswirklichkeit gearbeitet werden kann oder nicht (16, 17). Dies gilt jedoch für die Ergebnisse jeder Form von Studie, auch für angeblich so alltagsnahe Vorgehensweisen wie die Zusammenstellung von Fällen und die Auswertung von Registern. Die Frage der Alltagsrelevanz (respektive der externen Validität) ist kein Spezifikum von RCT, sie ist vielmehr vom Studiendesign grundsätzlich getrennt zu sehen. Im Übrigen ist aus den Auswirkungen vieler RCT auf die Behandlungswirklichkeit unmittelbar abzuleiten, dass deren Ergebnisse für den Alltag als relevant und entscheidungsleitend angesehen worden sind.

Die aufgeführten Kritikpunkte sprechen in ihrer Summe nicht überzeugend gegen RCT und zeigen vielmehr oft Informationsdefizite. Auffallend ist, dass sich viele Argumente, die gegen RCT und ihre Ergebnisse vorgebracht werden, nicht gegen die Randomisierung, sondern gegen den Studienansatz aussagefähiger, das heißt prospektiv geplanter, vergleichender Inter-

durch Patienten und die selektive Wahrnehmung von Ärzten, der Arztwechsel von unzufriedenen Patienten und damit ein unvollständiges Follow-up, eine verzerrte Erinnerung und anderes mehr tragen dazu bei, dass im Versorgungsalltag Therapieeffekte oft zu positiv eingeschätzt werden. Die Versorgungspraxis gibt also ein tendenziell geschöntes Feedback. Sicher spielt auch eine Rolle, dass wirkungslose, ja sogar gefährliche Untersuchungen und Behandlungen eine Beziehungsfunktion erfüllen können: die Bewältigung von Angst bei Patient und Arzt angesichts allgegenwärtiger Unsicherheit, Vermittlung von Kompetenz, Vertrauen und Hoffnung („Droge Arzt“).

Zudem sind in die Entwicklung innovativer Technologien viel Zeit und Geld investiert worden, wissenschaftliche Karrieren und die Interessen ganzer Berufsgruppen sind mit ihnen verknüpft. Ein RCT mit negativem Ergebnis ist deshalb in den Augen von Forschern und Entwicklern, Herstellern, Ärzten und auch hoffnungsvollen Patienten ein bedrohliches Risiko, das man nur zu gern zu umgehen sucht.

Aderlässe und Klistiere für jegliche Beschwerden, Bettruhe bei Rückenschmerzen, zu großzügig verordnete kardiale Antiarrhythmika, Schonung des Herzkranken – RCT waren und sind das aufklärerische Instrument, um Vorurteilen und gefährlichen Praktiken zu begegnen. Ärzte müssen nachweisen, dass die von ihnen vorgeschlagenen Behandlungen nachweislich mehr nutzen als schaden. Dies lässt sich nur mit wissenschaftlich validen Studiendesigns belegen, mit RCT an prominenter Stelle. Diese kritische Evaluation wird damit zu einem zentralen Definitionskriterium eines verantwortungsvollen therapeutischen Handelns.

Randomisierte kontrollierte Studien waren und sind das aufklärerische Instrument, um Vorurteilen und gefährlichen Praktiken zu begegnen.

Einsatz (und dem persönlichen Risiko) des Einzelnen jedenfalls einen absehbaren Nutzen gegenüberstellen zu können. Dies und damit die ethische Vertretbarkeit ist in qualitativ unzureichenden Studien per se nicht gegeben.

6. Die Ergebnisse aus RCT bildeten die Praxis („die Behandlungswirklichkeit“) nicht ab.

Es ist weder möglich noch von Interesse, eine Nutzenfragestellung unter Berücksichtigung aller Aspekte der „Behandlungswirklichkeit“ zu beantworten. Prospektive vergleichende Interventionsstudien sollen eine konkrete Frage so valide wie möglich beantworten. Sie müssen dazu andere Fragen und andere Aspekte ausblenden. Es ist im Einzelfall zu-

ventionsstudien als solche richten. Dies muss aber als Votum für die Verwendung qualitativ mangelhafter Studiendesigns für Fragestellungen verstanden werden, für die prospektiv vergleichende Studien ohne Zweifel das beste Studiendesign sind. Dies hätte auch ethische Implikationen.

Die „Erfindung“ des Prinzips der Randomisierung liegt schon mehr als 50 Jahre zurück, aber trotz der weitgehenden Akzeptanz dieses Goldstandards sind offenbar immer noch Diskussionsbeiträge wie dieser nötig. Warum ist dies so?

Die Resultate von RCT widersprechen oft der unmittelbaren klinischen Erfahrung. Der günstige Spontanverlauf von Erkrankungen, die selektive Symptomschilderung

■ Zitierweise dieses Beitrags:
Dtsch Arztebl 2008; 105(11):A 565–70

Anschrift für die Verfasser
Prof. Dr. med. Jürgen Windeler
Medizinischer Dienst der Spitzenverbände (MDS)
Lützowstraße 53, 45141 Essen

 **Literatur im Internet:**
www.aerzteblatt.de/lit18

LITERATURVERZEICHNIS HEFT 11/2008, ZU:

RANDOMISIERTE KLINISCHE STUDIEN

Kritische Evaluation ist ein Wesensmerkmal ärztlichen Handelns

Die gegen randomisierte klinische Studien vorgebrachten Einwände sind nicht überzeugend und zeigen oft Informationsdefizite

Jürgen Windeler, Gerd Antes, Johann Behrens,
Norbert Donner-Banzhoff, Monika Lelgemann

LITERATUR

1. Willich SN: Randomisierte kontrollierte Studien: Pragmatische Ansätze erforderlich. Dtsch Arztebl 2006; 103(39): A 2524.
2. Donner-Banzhoff N, Mayer-Berger W, Gelbrich G: Medikament-freisetzende versus konventionelle Stents – GERSHWIN-Studie zur Vermeidung von Koronar-Restenosen. Dtsch Arztebl 2006; 103(15): A 1019.
3. Niroomand F: Evidenzbasierte Medizin: Das Individuum bleibt auf der Strecke. Dtsch Arztebl 2004; 101(26): A 1870.
4. 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.
5. 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.
6. 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.
7. Harvey S, Young D, Brampton W, Cooper AB, Doig G, Sibbald W, Rowan K: Pulmonary artery catheters for adult patients in intensive care. Cochrane Database Syst Rev. 2006 Jul 19; 3: CD003408.
8. Rubin DB: The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Stat Med 2007; 26: 20–36.
9. Senn S: Testing for baseline balance in clinical trials. Stat Med 1994; 13: 1715–26.
10. Koller M, Lorenz W, Abel U: Methodenvielfalt in der klinischen Forschung. MMW 2006; 148: 85–91.
11. Donner-Banzhoff N, Lelgemann M: Ein neuer Maßstab – Aktuelle Studien verlangen veränderte Beurteilungskriterien. Z Ärztl Fortbild Qualitätssich 2003; 97: 301–6.
12. Behrens J: Einziger Goldstandard RCT? Gleiche Gütekriterien, unterschiedliche Validierungstechniken in „qualitativen“ und „quantitativen“ Interventions- und Evaluationsstudien. Gesundheitswesen 2002; 64.
13. van den Bent MJ, Afra D, de Witte O, Ben Hassel M, Schraub S, Hoang-Xuan K et al: Long-term efficacy of early versus delayed radiotherapy for low-grade astrocytoma and oligodendroglioma in adults: the EORTC 22845 randomised trial. Lancet 2005; 366: 985–90.
14. Demedts M, Behr J, Buhl R, Costabel U, Dekhuijzen R, Jansen HM et al: High-dose acetylcysteine in idiopathic pulmonary fibrosis. NEJM 2005; 353: 2229–42.
15. Windeler J, Lange S: Nutzenbewertung in besonderen Situationen – Seltene Erkrankungen. ZEFQ 2008; 102: 25–30.
16. Rothwell PM (Hrsg.): From randomised trials to personalised medicine. Elsevier 2007.
17. Behrens J, Langer G: Evidence-based Nursing and Caring. Bern, Oxford: Huber 2006.