

Armin Koch
Ulrich Abel

Abteilung Medizinische Biometrie, Universität
Heidelberg, Deutschland

Die Rolle der Randomisation in klinischen Studien

Schlüsselwörter

Klinische Studien · Randomisation · Beobachtungsstudien

Key Words

Clinical trials · Randomisation · Observational studies

Zusammenfassung

Randomisierte klinische Studien sind der akzeptierte Goldstandard für den Vergleich von Behandlungen. Nach wie vor bestehen jedoch Kontroversen in bezug auf die Durchführbarkeit, Adäquatheit und ethische Vertretbarkeit von randomisierten Studien. Gelegentlich werden der Randomisation Eigenschaften zugeschrieben, die nicht direkt aus der Zufallszuteilung der Therapien zu den Patienten resultiert. Dies mag aus historischen, pragmatischen oder erzieherischen Gründen verständlich sein, lässt sich jedoch aus wissenschaftstheoretischer Sicht nicht gut begründen. Dieser Artikel argumentiert, dass eine prinzipielle Notwendigkeit für die Durchführung nichtrandomisierter Studien besteht, klärt die Rolle der Randomisation in randomisierten klinischen Studien und prüft die Argumentation gegen die Aussagekraft beobachtender Studien. Es wird gezeigt, dass es bis jetzt keinen Grund gibt, anzunehmen, dass sorgfältig geplante und ausgewählte Beobachtungsstudien zu Resultaten führen, die sich von den Resultaten vergleichbarer randomisierter Studien unterscheiden. Obgleich man gut beraten ist, wann immer möglich randomisierte Studien durchzuführen, muss auch vermerkt werden, dass vielfach Beobachtungsstudien verbessert werden können, indem man darauf achtet, dass die Randomisation der einzige Unterschied zwischen einer Beobachtungsstudie und einer vergleichbaren randomisierten Studie ist.

Summary

The Place of Randomisation in Clinical Trials

Randomised clinical trials are nowadays the accepted gold standard for treatment comparisons. However, controversies still exist concerning the possibility to perform randomised trials or the adequacy and the ethical appropriateness of randomisation. Randomisation is sometimes credited with advantages it does not possess, and an extremely negative view or even a categorical rejection of non-randomised trials is found. This attitude may be comprehensible from a historical, pragmatic or educational viewpoint, but is not well-founded on epistemological grounds. This article argues that non-randomised clinical trials are needed, clarifies the role of randomisation in clinical trials and analyses the arguments raised against the validity of results from observational studies. It is shown that it has not been demonstrated up to now that well-designed and analysed observational studies would have yielded results that are distinct or even qualitatively different from results of similar randomised clinical trials. Although one is well advised to randomise whenever possible, there is still room for a considerable improvement of observational studies: Randomisation itself should be the only difference between observational studies and randomised clinical trials.

Einleitung

Seit ihrer Einführung in die Therapieforschung im Jahre 1948 hat die randomisierte klinische Studie, in der die zu vergleichenden Therapien den Studienpatienten nach einem Zufallsprinzip zugeordnet werden, eine beeindruckende Erfolgsgeschichte hinter sich gebracht. Heute ist sie der goldene Standard der Wirksamkeitsprüfung medizinischer Therapien und in der Regel die Voraussetzung

für die Zulassung von Arzneimitteln. Ob und wann randomisiert werden kann oder muss, ist jedoch nach wie vor der Gegenstand einer Kontroverse: Vertreter nichtrandomisierter Therapievergleiche stellen die Durchführbarkeit, Adäquatheit, ethische Vertretbarkeit und Aussagekraft randomisierter Studien in Frage. Hardliner unter den Kritikern nichtrandomisierter Studien vertreten die Auffassung, dass praktisch jede relevante Fragestellung im randomisierten Vergleich untersucht werden kann. Die Fronten können dabei,

wenn auch nicht scharf umrissen, so doch grob abgesteckt werden: Die Gegner der Randomisation finden sich gehäuft im Bereich der sogenannten «unkonventionellen Therapieverfahren» und in der Chirurgie. Die engagierten Befürworter finden sich bei den Internisten, Biometrikern und den Vertretern der Zulassungsbehörden.

Nach unserer Auffassung besteht nach wie vor die Notwendigkeit, in der Medizin für die Durchführung randomisierter klinischer Studien zu werben. Zu kritisieren ist jedoch, dass manche Autoren dabei der Randomisation Eigenschaften zuschreiben, die sie nicht hat, und dass die Randomisation zum Gütesiegel stilisiert wird, ohne das Studienergebnisse von vornherein als minderwertig erachtet werden. Da es die Aufgabe der Biometrie ist, das Methodenwerkzeug für den Erkenntnisgewinn in der Medizin kritisch zu bewerten, soll diese Arbeit klären, was genau die Randomisation garantiert, und die Argumente gegen nichtrandomisierte Therapievergleiche prüfen.

Die Rolle der Randomisation in randomisierten Studien

Rimm und Bortin haben bereits 1978 festgestellt, dass die randomisierte Studie nicht nur ein Ritual ist, sondern beinahe als Religion aufgefasst werden kann, in der an die Stelle einer Erörterung von Sinn und Nutzen der Randomisation das Gebot «Du sollst randomisieren» tritt [1]. Gelegentlich werden der Randomisation von Methodikern fast mystische Eigenschaften zugeschrieben. Die dann fehlende nüchterne Betrachtungsweise ist von Feinstein als «gläubige Ehrerbietung» verspottet worden [2]. So wird behauptet, dass die Randomisation (1) Strukturgleichheit (durch Ausschaltung der unterschiedlichen Formen von Selektionsverzerrung) gewährleiste, (2) notwendig oder hinreichend sei, einen beobachteten Effekt kausal zuzuordnen, (3) die Verblindung ermögliche, (4) die «Basis für statistische Tests» sei.

Keine dieser Aussagen ist strenggenommen korrekt: Ohne Zusatzvoraussetzungen (wie z. B. die Beobachtungsgleichheit) schafft die Randomisation keineswegs die Voraussetzung für statistische Tests. Sie ist auch nicht notwendig für die Anwendung statistischer Tests (wie etwa die durchaus zulässige Testung im Rahmen der statistischen Modellierung zeigt). Dass eine Verblindung (von Arzt, Patient, Beobachter) auch bei systematischer Zuordnung der Therapien möglich ist, ist evident. Ebenso führt die Randomisation auch nicht zu «Strukturgleichheit» (sonst wären Prozeduren wie eine stratifizierte Randomisation überflüssig). Die Randomisation ermöglicht lediglich Wahrscheinlichkeitsaussagen über Unterschiede zwischen den Therapiegruppen in bezug auf sämtliche zum Zeitpunkt der Randomisation festgestellten und prinzipiell feststellbaren Messgrößen der Patienten.

Randomisierten Studien werden häufig Vorzüge (wie z. B. die Beobachtungsgleichheit) zugeschrieben, die nicht aus der Zufallszuteilung der Therapien per se folgen, sondern daraus, dass sie zu den prospektiven Parallelgruppenstudien zählen, deren Prozedere und Fragestellung vor Versuchsbeginn in einem Studienprotokoll niedergelegt werden müssen und die einer Ethikkommission zur genauen Prüfung vorzulegen sind.

Die prinzipielle Notwendigkeit nichtrandomisierter Studien

Man muss zur Kenntnis nehmen, dass nicht alle Studien randomisiert durchgeführt werden können. Besonders in der Chirurgie scheint die Randomisation problematisch zu sein. So gibt es Untersuchungen, die belegen, dass nur eine Minderzahl der chirurgischen Interventionen in randomisierten Studien evaluiert wurden [3]. Des Weiteren hat der Anteil randomisierter Studien an allen klinischen Studien in der Chirurgie trotz der Erfolge des Konzepts in anderen Bereichen der Medizin nicht zugenommen [4]. Eine Untersuchung von 47 im New England «Journal of Medicine» publizierten Parallelgruppenstudien ergab, dass bei nur zwei der zwölf nichtrandomisierten Studien eine Randomisation möglich gewesen wäre [5].

Eine Randomisation wird unter anderem dann auf Probleme stoßen und praktisch aus durchaus nachvollziehbaren Gründen scheitern, wenn etwa die Patienten, deren Zustimmung erforderlich ist, eine klare Präferenz für eine der zu vergleichenden Therapien haben oder eine derartige Präferenz seitens der behandelnden Ärzte vorliegt. Darüber hinaus gibt es eine prinzipielle Unmöglichkeit, auf nichtrandomisierte Therapiestudien zu verzichten: Zu viele therapeutische Fragestellungen stehen zur Untersuchung an und zu viele Faktoren beeinflussen die Auswahl der geeigneten Therapie für den einzelnen Patienten [6].

Die Notwendigkeit, das Wissen aus randomisierten klinischen Studien mit Hilfe von Beobachtungsstudien zu erweitern, ist – als sogenannte «cross design synthesis» – im Rahmen von Richtlinien vom amerikanischen Kongress anerkannt und verabschiedet worden [7].

Die Biometrie muss sich mit der Aussagekraft nichtrandomisierter Studien beschäftigen, weil randomisierte Studien (1) nicht immer durchführbar oder angemessen sind, (2) zwar prinzipiell möglich sind, aber bisher nicht durchgeführt wurden, (3) neben randomisierten Vergleichen auch nichtrandomisierte Vergleiche zweier Behandlungsverfahren existieren.

Die Argumentation gegen nichtrandomisierte Studien zum Therapievergleich

Ein Verzicht auf die Randomisation ist wissenschaftlich und wissenschaftspolitisch unklug, eben weil die Therapieforschung nicht nur das Ziel hat, die bessere von z. B. zwei Therapien zu identifizieren, sondern auch dieses Ergebnis in einer Form zu gewinnen, die für andere nachvollziehbar ist, so dass sie sich den getroffenen Schlussfolgerungen nicht entziehen können. Ein unnötiger Verzicht auf die Randomisation führt auch dazu, dass für einen beobachteten Effekt ausser der eigentlich interessierenden Möglichkeit, dass eben z. B. die Therapie A der Therapie B überlegen ist, eine Reihe von anderen «Ursachen» in Betracht gezogen werden müssen. Mögliche Verzerrungsquellen in nichtrandomisierten Therapievergleichen werden in einer Fülle von Arbeiten diskutiert [8]. Tabelle 1 gibt eine unserer Auffassung nach vollständige Liste die-

Tab. 1. Ursachen für einen beobachteten «Effekt» (Unterschied zwischen zwei Behandlungsgruppen in der Zielvariablen) in einer vergleichenden Studie

Therapieeffekte (unterschiedliche Wirksamkeit)
Zufallseffekte der Stichprobenauswahl
Unterschiede in den Patientencharakteristika bei Studienbeginn
Definition der Krankheit (Nomenklatur)
Diagnostik (Stadienwanderung, «zero-point shift»)
Ein- und Ausschlusskriterien (z. B. Ausschluss prognostisch ungünstig gelagerter Fälle)
Selbstselektion durch die Patienten (z. B. Präferenz gewisser Patientengruppen für eine der Therapien)
Herkunft und Zuweisung der Patienten zur Prüfinstitution
Qualität der Therapiedurchführung und Engagement der Ärzte
Unterschiede in der Patientenmotivation
Unterschiede in der allgemeinen Patientenversorgung und im Umfeld
Begleitmedikation und supportive Massnahmen
Lifestyle
Umfeld (familiäres und berufliches Umfeld, Umwelteinflüsse)
Beobachtungsungleichheit
Definition des Therapieerfolges
Messung des Therapieerfolges
Qualität der Datenerhebung und der Follow-up
(Vollständigkeit, Zuverlässigkeit)

ser Ursachen an. In diesem Zusammenhang ist es aufschlussreich zu überlegen, welche dieser Verzerrungsquellen tatsächlich durch die Randomisation per se und welche Verzerrungen durch andere wünschenswerte Eigenschaften (wie z. B. die Durchführung einer prospektiven Studie), die eben mit der Randomisation einhergehen, ausgeschaltet werden.

Ein wichtiger Schritt in der Auseinandersetzung mit nichtrandomisierten Studien ist es, neben der Art der Verzerrung auch Anhaltspunkte für deren Ausmass in Abhängigkeit vom gewählten Studiendesign und der Qualität der Studie im allgemeinen zu erhalten.

Die Argumentation gegen nichtrandomisierte Studien stützt sich im wesentlichen auf zwei Kategorien von empirischen Untersuchungen. Zum einen sammeln Biometriker zu didaktischen Zwecken sogenannte «horror stories» (anekdotische Beispiele), mit deren Hilfe man hervorragend argumentieren kann, wenn es gilt, klinische Kooperationspartner von der Notwendigkeit der Randomisation zu überzeugen. Berühmte Beispiele, wie die Shunt-Operation bei Leberzirrhosen mit Ösophagusvarizen [9], Vitamin C bei fortgeschrittenen Krebserkrankungen [10] und die Radiatio bei Schilddrüsentumoren [11], scheinen durch den Vergleich der Resultate aus nichtrandomisierten Studien mit nachfolgenden randomisierten Therapiestudien, die die vorausgegangenen Ergebnisse nicht bestätigen konnten, die Gefahren eines Verzichts auf die Randomisation mehr als eindeutig zu belegen. So beeindruckend diese Beispiele auch sind, werden sie dem gestellten Problem aus mehreren Gründen nicht gerecht:

(1) Es ist nicht auszuschliessen, dass «horror stories» eine Negativselektion darstellen: Folgt auf eine Studie mit historischen Kontrollen oder eine Registerauswertung eine randomisierte Studie, die das vorausgegangene Ergebnis bestätigt, so dürfte dies kaum Eingang in die methodische Literatur finden.

(2) Eventuell liegt ein Publikationsbias vor: «nichtsignifikante» Resultate aus Beobachtungsstudien werden, weil eine wichtige Funktion von Beobachtungsstudien die Hypothesengenerierung für die weitere Forschung ist, wahrscheinlich weniger das Interesse der Editoren von wissenschaftlichen Fachzeitschriften finden als entsprechende Resultate aus randomisierten Studien und folglich schwieriger zu publizieren sein. Durch die im Rahmen der Studienplanung erforderliche Fallzahlschätzung führt der Anspruch der Randomisation in der Regel zu grösseren und häufiger zu multizentrischen Erhebungen, in denen dann auch ein grösserer Publikationsdruck besteht.

(3) Fast immer handelt es sich bei den angeführten Beispielen um den Vergleich der Resultate aus randomisierten Studien und aus Studien mit historischen Kontrollen. Vollzieht sich jedoch die Therapieforschung nach gängigen Regeln, so stehen die nichtrandomisierten Vergleiche, die – wie schon erwähnt – zur Hypothesengenerierung herangezogen werden, am Anfang der Entwicklung, auf die dann die randomisierten Studien folgen. Hinzu kommt, dass die Kontrollpatienten zeitlich noch früher behandelt wurden als die Patienten der Therapiegruppe der historisch kontrollierten Studie. Wird basierend auf einem historischen Vergleich ein Therapieeffekt angegeben, so wird dabei die in der Regel unplausible Annahme gemacht, dass ausser der betrachteten Therapie keine wesentlichen Veränderungen in bezug auf die Diagnose der Erkrankung oder die Patientenversorgung stattgefunden haben. Ein Indiz dafür, dass die zu optimistischen Therapieeffekte in Studien mit historischen Kontrollen eben gerade eine Folge des extrem schlechten Abschneidens der Kontrollgruppe aufgrund von nichterfassten oder nichterfassbaren Veränderungen in der Patientenversorgung sind, ergibt sich bei genauerer Betrachtung der klassischen Gegenüberstellung von Studien mit historischen Kontrollen und randomisierten Studien von Sacks et al. [9]. Es zeigt sich, dass die zeitlich näher an den randomisierten Studien liegenden Therapiearme der Studien mit historischen Kontrollen ein vergleichbares Ergebnis erzielt haben.

Somit sind die obengenannten Arbeiten geeignet, die Ergebnisse von Studien mit historischen Kontrollen oder retrospektive Registerauswertungen fragwürdig erscheinen zu lassen. Sie können jedoch nicht als Argumente gegen gut geplante prospektive nichtrandomisierte Studien verwendet werden.

Die zweite Kategorie von empirischen Untersuchungen sind sogenannte synthetische historische Vergleiche, bei denen etwa in einer randomisierten Studie Schätzungen des Therapieeffektes auf der Basis von Patienten des einen Therapiearms aus einer frühen Studienphase mit den Patienten des anderen Therapiearms in einer späteren Studienphase mit den entsprechenden Schätzungen im gleichen Zeitsegment verglichen werden. Byar et al. [12] untersuchen dies an einer randomisierten Studie zum Vergleich von Östrogentherapie und Placebo beim Prostatakarzinom. Hier hat einzig der synthetische historische Vergleich zu einem signifikanten Therapieeffekt geführt, der allerdings nach Adjustierung für die Variablen Patientenalter und Allgemeinzustand bei Aufnahme verschwand. Auch dieser Fall belegt die Gefahren von Auswertungen, die auf historischen Kontrollen basieren oder kann verwendet

werden, um gegen den Unsinn unadjustierter Vergleiche in Beobachtungsstudien zu argumentieren. Zur Argumentation gegen nichtrandomisierte prospektive Therapiestudien kann er jedoch nicht herangezogen werden.

Fazit

Es ist nicht das Ziel dieser Arbeit, für die Durchführung nichtrandomisierter Studien zu plädieren. Ganz im Gegenteil besteht nach unserer Auffassung Begründungspflicht, wenn man von diesem Goldstandard abweichen will. Andererseits ist die bei Methodikern häufig anzutreffende extrem ablehnende Haltung gegenüber

nichtrandomisierten Studien schlecht begründet. Weder die abschreckenden Beispiele aus der medizinischen Literatur noch die empirischen Untersuchungen beantworten die eigentliche Frage, ob gut geplante Beobachtungsstudien nicht im Regelfall zu ähnlichen Ergebnissen wie entsprechende randomisierte Studien gelangt wären. Die ablehnende Haltung mancher Methodiker ist auch schädlich, da diese nicht mehr darauf achten, dass all die übrigen Kriterien (wie z. B. Beobachtungsgleichheit usw.), die zum Erfolg des Konzeptes «randomisierte klinische Studie» beitragen, adäquate Berücksichtigung finden und folglich nichtrandomisierte Studien von nicht so guter Qualität sind, wie dies eigentlich der Fall sein müsste.

Literatur

- 1 Rimm AA, Bortin M: Clinical trials as a religion. *Biomedicine (special issue)* 1978;28:60–63.
- 2 Feinstein AR: An additional basic science for clinical medicine: II. The limitations of randomized trials. *Ann Intern Med* 1983;99:544–550.
- 3 Miller JN, Colditz GA, Mosteller F: How study design affects outcomes in comparisons of therapy. II: Surgical. *Stat Med* 1989;8:455–466.
- 4 Solomon MJ, Leod RS: Clinical studies in surgical journals – we have improved? *Dis Colon Rectum* 1993;36:43–48.
- 5 Lavori PW, Louis TA, Bailar JC, Polanski M: Design for experiments – Parallel comparisons of treatment. *N Engl J Med* 1983;309:1291–1298.
- 6 Dupont WD: Randomized vs. historical clinical trials: Are the benefits worth the costs? *Am J Epidemiol* 1985;122:940–946.
- 7 Anonymous: Cross design synthesis: A new strategy for studying medical outcomes? *Lancet* 1992;340:944–946.
- 8 Abel U, Koch A: Randomisation in klinischen Studien: Empirisch begründet oder nur ein Dogma? *Internist* 1997;38:318–324.
- 9 Sacks H, Chalmers TC, Smith J: Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233–240.
- 10 Moertel CG, Fleming TR, Creagan ET, Rubin J, O’Connell MJ, Ames MM: High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy. A randomized double-blind comparison. *N Engl J Med* 1985;312:137–141.
- 11 Byar DP: The use of data bases and historical controls in treatment comparisons. *Recent Results Cancer Res* 1988;111:95–98.
- 12 Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, Demets DL, Ellenberg JH, Gail MH, Ware JH: Randomized clinical trials. *N Engl J Med* 1976;295:74–80.