Serie: Methodik

Wien Med Wochenschr (2008) 158/7–8: 234–239
DOI 10.1007/s10354-008-0526-y

WMW
Wiener Medizinische Wochenschrift
Printed in Austria

Randomisiert kontrollierte Studien

Philipp Mad, Rosemarie Felder-Puig und **Gerald Gartlehner**

Ludwig Boltzmann Institut für Health Technology Assessment, 1090 Wien, Österreich

Eingegangen am 9. August 2007, angenommen nach Revision am 10. Jänner 2008

Randomised Controlled Trials

Summary. Nowadays the Randomised Controlled Trial (RCT) is seen as the gold standard for estimating the effectiveness of an observed intervention, achieving the highest hierarchy of evidence of primary research settings. Its study design basically includes two groups of patients, an intervention group and a control group; patients are randomly allocated to these two groups. After intervention or control intervention took place, predefined outcomes are quantified and compared in the two groups. The study design aims at eliminating all confounding and distorting factors (Bias and Confounder), so that different outcomes between the groups can be only explained by the intervention.

There is a broad variation of quality of published RCTs. The reliability of results and extent to which findings provide a correct basis for generalisation to other circumstances needs to be validated. As part of a methods series of the Wiener Medizinische Wochenschrift this paper will discuss principles of study design, critical appraisal and limitations of RCTs.

Key words: Randomised controlled trial, methodology, study design.

Zusammenfassung. Die randomisiert kontrollierte Studie (engl. Randomised Controlled Trial, RCT) steht in der Hierarchie der Evidenz an der obersten Stufe aller Primärstudien, sie ist heute der Goldstandard, um den Effekt von Interventionen nachzuweisen. Das Studiendesign eines RCT besteht im Wesentlichen aus einer Interventions- und einer Kontrollgruppe, zu denen Patienten per Zufall zugeordnet werden. Nach Intervention bzw. Kontrollintervention wird das Auftreten vordefinierter Interventionsziele (Endpunkte) gemessen. Das Ziel des aufwändigen Studiendesigns ist es, alle störenden und verzerrenden Einflüsse auf die Studienendpunkte zu eliminieren, sodass ein gemessener Unterschied zwischen den Gruppen nur durch die Intervention erklärbar wird.

Die methodische Qualität veröffentlichter RCTs ist sehr unterschiedlich. Um beurteilen zu können, in wieweit die Studienergebnisse einerseits die Realität abbilden und andererseits für die eigene Situation verwertbar sind, müssen Aspekte der internen und externen Validität der Studie betrachtet werden. Als Teil der Methoden Serie der Wiener Medizinischen Wochenschrift wird in diesem Artikel auf die Grundprinzipien von Studiendesign, kritischer Beurteilung und Limitationen von RCTs eingegangen.

Schlüsselwörter: Randomisiert kontrollierte Studie, Methodik, Studiendesign.

Korrespondenz: Dr. Philipp Mad, Ludwig Boltzmann Institut für Health Technology Assessment,Garnisongasse 7/20, 1090 Wien, Österreich.
Fax: ++43-1-236811999
E-Mail: philipp.mad@hta.lbg.ac.at

Einleitung

In der medizinischen Praxis werden verschiedenste Strategien angewandt, um Patienten erfolgreich zu behandeln. Diese *Interventionen* können sehr vielgestaltig sein und umfassen medikamentöse und physikalische Therapien, chirurgische Eingriffe, diagnostische Vorgänge, präventive Maßnahmen, u.v.m. Ebenso vielgestaltig sind die *Ziele* dieser Interventionen, wie eine Verlängerung der Lebenserwartung, Verkürzung des Spitalsaufenhaltes, Linderung von Beschwerden, Vermeidung von Nebenwirkungen oder die Verhinderung des Krankheitsausbruches. Möchte man wissen, wie gut eine Intervention geeignet ist, um ein erwünschtes Behandlungsziel zu erreichen, so stellt man die Frage nach deren *Wirksamkeit*.

Um die Wirksamkeit medizinischer Interventionen nachzuweisen, müssen bestimmte Vorraussetzungen gegeben sein: erstens muss eine Gruppe von Patienten die Intervention erhalten (Interventionsgruppe), zweitens muss diese Gruppe mit einer weiteren Gruppe von Patienten verglichen werden, welche diese Intervention nicht erhalten (Kontrollgruppe). Drittens dürfen sich Interventionsgruppe und Kontrollgruppe mit Ausnahme der Intervention (Risikofaktor) nicht systematisch unterscheiden, denn würden neben der Intervention noch andere Unterschiede zwischen den beiden Gruppen bestehen, so könnte der beobachtete Effekt in der Interventionsgruppe auch durch diese anderen Unterschiede erklärt werden [1]. Um dies zu erzielen, müssen die Patienten, die für die untersuchte Intervention in Frage kommen, nach dem Zufalls-

prinzip der Behandlungs- oder Kontrollgruppe zugeordnet werden (Randomisierung).

Die beste Methode, um die Wirksamkeit medizinischer Interventionen unter den angeführten Voraussetzungen nachzuweisen, wird heute als „randomisiert kontrollierte Studie“ engl. *Randomised Controlled Trial* (RCT) bezeichnet [2]. Aus Sicht der evidenzbasierten Medizin haben RCTs unter allen Primärstudien den höchsten Grad an Aussagekraft, denn nur durch sie kann ein Zusammenhang (Kausalität) zwischen Intervention (Risikofaktor) und Behandlungsziel (Endpunkt) hergestellt werden. RCTs werden heute als der „Goldstandard“ zur Bestimmung der Effektivität von Interventionen angesehen [3].

Der folgende Artikel hat zum Ziel, die bei RCTs angewendete Methodik sowie die Interpretation der Ergebnisse für die Entscheidungsfindung im medizinischen Alltag näher zu bringen.

Von der Fragestellung zum Studiendesign: die Hierarchie der Evidenz

Vereinfacht dargestellt befassen sich die meisten klinischen Studien mit *Risikofaktoren* und *Endpunkten*. Unter dem Begriff Risikofaktor versteht man bestimmte Eigenschaften, die Krankheitszustände auslösen oder verschlimmern, oder aber auch vor Krankheitszuständen schützen können. So können beispielsweise hoher Blutdruck, Lebensalter, Adipositas oder Verhaltensweisen wie Zigarettenrauchen als Risikofaktor gesehen werden; bei Interventionsstudien gelten Behandlungen wie Medikamente und chirurgische Interventionen als „Risikofaktor“, um bestimmte Krankheitszustände zu verhindern, zu heilen oder zu bessern.

Der Endpunkt ist ebenfalls eine Eigenschaft, die meist einen Krankheitszustand beschreibt. Häufig in Studien verwendete Endpunkte sind der Tod (Mortalität) oder das Auftreten eines bestimmten Krankheitsbildes (Morbidität), aber auch der Zeitpunkt der Wiederaufnahme in das Krankenhaus oder die Abwesenheit von Krankheitsbildern über einen bestimmten Zeitintervall.

In Studien können die Häufigkeit und die Verteilung von Risikofaktoren, die Häufigkeit und Verteilung von Endpunkten, sowie die Zusammenhänge zwischen Risikofaktoren und Endpunkten untersucht werden, wobei die Art der Fragestellung das Studiendesign bestimmt: in Querschnittsstudien wird die Häufigkeit (Prävalenz) von Risikofaktoren und Endpunkten gleichzeitig gemessen, um beispielsweise die Größenordnung bestimmter Krankheiten in einer Population zu erfassen. Mittels Fall-Kontroll Studie wird die unterschiedliche Verteilung des Risikofaktors in einer Gruppe von Individuen mit Endpunkt und einer willkürlich ausgewählten Kontrollgruppe retrospektiv bestimmt, bei Kohortenstudien wird der Risikofaktor in der Kohorte gemessen und danach das Auftreten des Endpunktes beobachtet, um einen zeitlichen Zusammenhang zwischen Risikofaktor und Endpunkt herzustellen. Um jedoch eine Kausalität zwischen Risikofaktor und Endpunkt bestmöglich darzustellen, bedarf es eines randomisiert kontrollierten Studiendesigns: Probanden werden rekrutiert, danach wird die Intervention nach dem Zufallsprinzip zugeteilt. Nach einer Beobachtung über einen vordefinierten Zeitraum (Studiendauer) kann dann der Endpunkt erfasst und eine Kausalität zu dem Risikofaktor bewiesen oder ausgeschlossen werden [1].

Um beispielsweise einen Zusammenhang zwischen Bluthochdruck (Risikofaktor) und Schlaganfall (Endpunkt) zu *beschreiben*, können daher in einer Querschnittstudie die Häufigkeit (Prävalenz) von Bluthochdruck und Schlaganfall in einem Patientenkollektiv dargestellt werden, in einer Fall-Kontroll-Studie würde bei Patienten mit Schlaganfall (Fälle) und einer ausgewählten Gruppe von Individuen ohne Schlaganfall (Kontrollen) die Häufigkeit von Bluthochdruck bestimmt werden. Um *nachzuweisen*, dass Bluthochdruck unbehandelt zu Schlaganfällen führen kann, müsste man Bluthochdruck-Patienten nach dem Zufallsprinzip in eine antihypertensive Behandlungsgruppe und eine Kontrollgruppe ohne antihypertensive Behandlung aufteilen, und danach das Auftreten eines Schlaganfalles (Endpunkt) in beiden Gruppen vergleichen.

Daraus wird ersichtlich, dass bestimmte Studiendesigns besser geeignet sind als andere, um die Frage nach der Effektivität einer bestimmten Behandlung zu beantworten [4], man spricht heute von einer „Hierarchie der Evidenz“ [5] der verschiedenen Studiendesigns (siehe *Tabelle 1*). An oberster Stelle der Primärstudien steht hier der RCT, durch dessen aufwändiges Studiendesign versucht wird, möglichst alle das Studienergebnis verändernden systematischen Fehler (Bias und Confounder, Begriffserklärung siehe Glossar der evidenzbasierten Medizin in *Tabelle 2*) auszuschalten oder zu minimieren. Allerdings konnte gezeigt werden, dass auch gut durchgeführte Beobachtungsstudien zu mit RCTs vergleichbaren Ergebnissen führen können [6]. Will man allerdings einen Wirksamkeitsnachweis durchführen, so sollte man unbedingt das Design eines RCT wählen.

Methodik und Durchführung randomisiert kontrollierter Studien

Im Folgenden wird in groben Zügen erläutert, wie ein RCT idealer Weise durchgeführt werden sollte. Die ein-

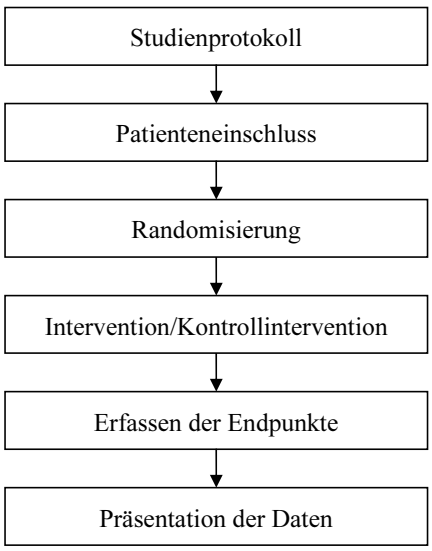

Abbildung 1: Ablauf einer randomisiert kontrollierten Studie

Tabelle 1: Hierarchie der Evidenz [5]

I.	Systematische Übersichtsarbeiten und Meta-Analysen mehrerer randomisiert kontrollierter Studien (Sekundärstudie)
II.	Randomisiert kontrollierte Studie
III.	Nicht-randomisierte, kontrollierte Studie (Kohortenstudie) bzw. Fall-Kontroll Studie
IV.	Nicht-kontrollierte Studie (Fallserie)
V.	Expertenmeinung

Tabelle 2: Glossar zur Evidenz-basierten Medizin [21]

Bias (Verzerrung): Tendenz der Studienergebnisse systematisch von den „wahren" Ergebnissen abzuweichen. Bias führt entweder zu einer Über- oder Unterschätzung der wahren Wirkung einer Maßnahme oder Exposition. Die Ursache dafür liegt vor allem in Design und der Durchführung der Studie und führen zu systematischen Unterschieden zwischen den Vergleichsgruppen.
Confounding (Störfaktor): Confounding liegt vor, wenn ein Faktor, der nicht direkt Gegenstand der Untersuchung ist, sowohl mit der Intervention / Exposition als auch mit der Zielgröße assoziiert ist und dadurch bei Aussagen über die Beziehung zwischen Intervention / Exposition und Zielgröße „Verwirrung" stiftet.
Evidenz: Der Begriff „Evidenz" leitet sich vom englischen Wort „evidence = Nachweis, Beweis" ab und bezieht sich auf die Informationen aus klinischen Studien, die einen Scahverhalt erhärten oder widerlegen.
Externe Validität (Übertragbarkeit, Anwendbarkeit): Beschreibt die Übertragbarkeit von Studienergebnissen auf die Patienten in der Routineversorgung, d.h. auf Patienten, die nicht an der Studie teilgenommen haben.
Interne Validität: Bezeichnet das Ausmaß, mit dem die Ergebnisse einer Studie die „wahren" Effekte einer Intervention/Exposition wiedergeben, d.h. frei von systematischen Fehlern sind. Die interne Validität beruht auf der Integrität des Studiendesigns und ist Voraussetzung für die Anwendbarkeit der Studienergebnisse in der Routineversorgung.

zelnen Arbeitsschritte eines RCTs sind in *Abbildung 1* vereinfacht zusammengefasst.

Patienteneinschluss und Randomisierung

Der Patienteneinschluss erfolgt gemäß der vordefinierten Ein- und Ausschlusskriterien. Prinzipiell gilt für RCTs, dass ein möglichst breites Spektrum an Patienten eingeschlossen werden sollte, da dies die Generalisierbarkeit des Studienergebnisses steigert (siehe Kapitel 5, externe Validität). Die Nachvollziehbarkeit des Patienteneinschlusses anhand dieser Kriterien ist wichtig, um die in der Studie gefundenen Ergebnisse für andere Patienten verallgemeinern zu können.

Danach erfolgt die Randomisierung, d.h. die zufällige Zuteilung der Patienten in eine Interventions- oder Kontrollgruppe. Randomisierung erfolgt heute weitgehend durch Computer-generierte Sequenzen. Der Sinn der Randomisierung liegt darin, dass bekannte und unbekannte Störfaktoren (Confounder) in beiden Gruppen gleich verteilt werden. Bei selektiver Patientenzuteilung können Confounder ungleich verteilt auftreten und dadurch das Ergebnis beeinflussen, durch eine gleichmäßige Verteilung der Confounder auf beide Gruppen kann dieser Effekt ausgeschalten werden. Voraussetzung dabei ist allerdings eine ausreichend große Stichprobe. Bei kleinen RCTs mit weniger als 300 Teilnehmern kann es durch Zufall, auch bei guter Randomisierung, zur ungleichmäßigen Verteilung von Confoundern kommen (Zufallsvariabilität). Bei großen RCTs sind Confounder gleichmäßig verteilt und ein Unterschied in den Ergebnissen (Endpunkten) zwischen den Gruppen kann nur mehr durch die Intervention erklärbar gemacht werden.

Die „zufällige" Gruppenzuteilung sollte allerdings nicht vorhersehbar sein. Dies wird als „*allocation concealment*" bezeichnet und bedeutet, dass es für Prüfärzte nicht nachvollziehbar sein soll, welcher Gruppe der nächste Patient zugeteilt wird. In der Praxis wird dies durch nummerierte, blickdichte Umschläge erzielt, die die Zuordnung für einzelne Patienten enthalten, durch automatische Telefonrandomisierung oder durch die Verwaltung der Randomisierungssequenz durch dritte, an der Studie unbeteiligte Personen.

Eine abwechselnde Zuteilung von Patienten oder eine Zuteilung nach Geburtsdatum wäre beispielweise vorhersehbar und daher nicht zulässig, weil sie zu Selektionsbias führen könnten.

Dass Randomisierung Sinn macht, zeigt folgende Studie: Bei insgesamt 45 klinischen Fragestellungen, zu denen sowohl randomisierte als auch nicht-randomisierte Studien durchgeführt wurden, wurden deren Studienergebnisse verglichen. Es konnte gezeigt werden, dass bei gleicher Fragestellung der Behandlungseffekt in nicht randomisierten Studien deutlich größer geschätzt wurde als in randomisierten Studien [7]. Randomisierung schützt daher vor einer Überschätzung des gemessenen Effektes, mit anderen Worten kann man ohne Randomisierung nicht sicher sein, ob der gemessene Effekt tatsächlich nur durch die Intervention erklärbar ist.

Verblindung

Nach der Randomisierung erfolgt die eigentliche Studienphase, in der die Intervention bzw. die Kontrollintervention durchgeführt wird. Die Kontrollintervention kann die Standard-Intervention, keine Intervention oder, vor allem bei Medikamentenstudien, ein Scheinmedikament gleichen Aussehens ohne pharmakologischen Wirkstoff (Plazebo) sein. Diese Phase sollte nach Möglichkeit verblindet durchgeführt werden. Im Idealfall wissen weder Patient noch behandelnder Arzt noch Studienbetreuer, ob der individuelle Patient der Interventions- oder Kontrollgruppe angehört, denn diese Information könnte die Beteiligten in ihrer Entscheidung und Bewertung beeinflussen, was wiederum das Studienergebnis verfälschen könnte [8, 9]. Oft kann man auf Grund der Intervention nicht alle an der Studie beteiligten Personen ausreichend verblinden, beispielweise weil der behandelnde Operateur natürlich weiß, welche OP-Technik er eingesetzt hat. Jedoch sollte bei Nicht-Verblindung dies (a) explizit erwähnt werden und (b) erklärt werden, warum eine Verblindung nicht möglich war. In Untersuchungen konnte festgestellt werden, dass in vielen RCTs nicht ausrei-

chend verblindet wurde, obwohl dies möglich gewesen wäre [10]. Verblindung ist vor allem dann wesentlich, wenn eine subjektive Beurteilung das Ergebnis beeinflussen kann. Beispielsweise kann alleine die Tatsache, ein neues Medikament gegen Schmerzen zu bekommen, zu Verzerrungen bei der Bewertung in einer Schmerzskala führen, weil „neu" häufig mit „besser" gleichgesetzt wird. Man nennt diese Verzerrung Measurement Bias. Bei objektiven, nicht beeinflussbaren Endpunkten, wie zum Beispiel der Gesamtmortalität, ist eine Verblindung nicht notwendig.

"Intention to treat" vs. "per protocol"

Nicht alle Patienten, welche in die Studie zu Anfang eingeschlossen wurden, durchlaufen und beenden diese auch so, wie das im Studienprotokoll vorgesehen ist. So haben z.B einige Patienten doch nicht die vorgesehene Therapie erhalten, wurden der falschen Gruppe zugeordnet oder haben die Studienteilnahme vorzeitig abgebrochen. Würden solche Patienten nun von der Analyse ausgeschlossen werden, so könnte man nicht mehr sicher sein, dass die unbekannten Störfaktoren, welche durch die Randomisierung auf beide Gruppen gleich verteilt wurden, auch weiterhin gleich verteilt blieben [11]. Um das zu verhindern, sollte die Datenauswertung gemäß einer „*Intention to treat*" Analyse durchgeführt werden [12]. Dies bedeutet, dass *alle* Patienten, die der Interventions- bzw. Kontrollgruppe zugeteilt waren, auch so in die Datenanalyse einbezogen werden, als ob sie die vorgesehene Intervention erhalten hätten, egal ob sie die Studie beenden, das Studienprotokoll befolgten, die richtige Intervention bekamen, oder nicht. Die Ergebnisse einer „Intention to treat" Analyse sind jedoch meist abgeschwächt im Vergleich zu einer *„Per protocol"* Analyse, bei der die Patienten nach der tatsächlich erhaltenen Therapie ausgewertet werden, unabhängig davon, welcher Gruppe sie zu Studienbeginn zugeteilt waren. Die „Per protocol" Analyse entspricht also eher den Idealbedingungen. Da aber auch in der täglichen Praxis nicht immer alles fehlerfrei abläuft, spiegelt die „Intention to treat" Analyse eher die in der Realität zu erwartenden Ergebnisse wieder, daher sind die „Intention to treat" Ergebnisse besser generalisierbar. In jedem RCT sollten (a) Informationen über jene Patienten enthalten sein, welche die Studie nicht beendeten, und (b) beschrieben werden, wie diese Patienten in die Berechnung der Effektgröße einbezogen wurden.

Kritische Evaluierung von RCTs

Obwohl der RCT als die best mögliche Methodik zur Bestimmung der Effektivität einer Intervention gilt, bedeutet dies nicht, dass die Ergebnisse aller als RCT bezeichneten Studien unkritisch übernommen werden können; vielmehr sind deutliche Unterschiede in der Studienqualität zu beobachten. Einerseits stellt sich die Frage, mit welcher methodischen Qualität die Studie durchgeführt wurde, diese Frage kann man mit der Beurteilung der *internen Validität* beantworten. Unter interner Validität versteht man das Ausmaß an Bemühungen, systematische Fehler (Bias) durch methodisch adäquates Studiendesign und Durchführung der Studie zu minimieren [13]. Bei hoher interner Validität kann davon ausgegangen werden, dass die gefundenen Ergebnisse der gesuchten „Wahrheit" nahe kommen.

Andererseits können auch valide, der Wahrheit sehr nahe kommende Studienergebnisse für die eigene Situation völlig unbrauchbar sein. Beispielsweise haben Studien, welche an Universitätsspitälern durchgeführt werden, oft nur begrenzt Aussagekraft für die Primärversorgung, da sich die Zusammensetzung von Patienten und behandelndem Personal sowie die Infrastruktur in diesen verschiedenen Versorgungssituationen deutlich unterscheiden. Die Frage nach der Generalisierbarkeit der Studienergebnisse wird als *externe Validität* bezeichnet [13]. Es gibt allgemein anerkannte Methoden, mit denen die interne und externe Validität einer Arbeit beurteilt werden können [14-16]. Hierbei werden einzelne Komponenten des RCTs analysiert und bewertet:

Fragen zur Beurteilung der Internen Validität

- Wie wurde randomisiert? Ist die Unvorhersehbarkeit der Gruppenzuordnung (allocation concealment) gewährleistet? Adäquate Methoden der Randomisierung wären: Computer-generierte Zufallszahlen, Randomisierung durch Dritte; vorhersehbare und daher schlechte Randomisierung wäre: Geburtsdatum, Wochentag des Einschlusses, alternierende Zuordnung, etc.
- Wurden alle Teilnehmer verblindet? Idealerweise sollten alle Beteiligten verblindet sein (Patienten, behandelnde Ärzte, Studienbetreuer), oft ist das aber nicht möglich (z.B. bei chirurgischen Interventionen). In jedem Fall sollten jene Personen verblindet sein, welche die Ergebnisse bewerten.
- Wurden Studienabbrecher/Drop Out Raten dokumentiert, und wie viele waren es? Drop-out Raten sollten immer angeführt sein; Arbeiten, die Patientenausschlüsse dokumentieren sind meist von höherer Qualität als jene ohne derartige Dokumentation [10]. Wenn viele Probanden während der Studie verloren gehen, dann kann ein zugrundeliegender systematischer Fehler (Selektions Bias) nicht mehr ausgeschlossen werden. Es gibt hierfür keine einheitlichen Grenzen, oft werden aber Drop-Out Raten von über 20 % als Qualitätsmanko angesehen.

Fragen zur Beurteilung der externen Validität

- An welcher Population wurde die Studie durchgeführt? Sind die Ein- und Ausschlusskriterien zu selektiv, dann sinkt die Generalisierbarkeit der gefundenen Ergebnisse. Wie oben angeführt können sich Studienpopulationen der Universitätsspitäler deutlich von Patienten der Primärversorgung unterscheiden.
- Klinische Relevanz: Entsprechen die in der Studie durchgeführten Interventionen der klinischen Realität, und waren die Endpunkte für Patienten relevant?
- Wurden Nebenwirkungen der Intervention dokumentiert? Es sollten alle Patienten befragt werden, ob durch die Intervention zu erwartende Nebenwirkungen während des Untersuchungszeitraumes eintraten, Ergebnisse dieser Befragung sollten in der Publikation enthalten sein.

- Wurden alle Patienten, die in die Studie eingeschlossen wurden, in der Analyse berücksichtigt („Intention-to-treat"), oder nur jene, welche die Studie beendet haben („per protocol")?
- Wurde die Größe der Fallzahl berechnet und dokumentiert? Zu kleine Fallzahlen können dazu führen, dass auch deutliche Unterschiede kein Signifikanzniveau erreichen, zu große Fallzahlen können auch kleine, klinisch irrelevante Unterschiede statistisch signifikant erscheinen lassen.

Grenzen von RCTs

Zufallsvariabilität

Selbst bei methodisch adäquat durchgeführten Studien mit hoher interner Validität kann man nicht zu 100 % sicher sein, dass das Studienergebnis der Wahrheit entspricht, denn das gemessene Ergebnis könnte auch nur durch Zufall zustande gekommen sein. In der medizinischen Biostatistik muss man mit dieser zufälligen Variabilität der Studienergebnisse leben, man kann aber eine gewisse „Schadensbegrenzung" betreiben, indem man die Wahrscheinlichkeit eingrenzt, mit welcher diese Zufallsvariabilitäten auftreten können. Hierbei werden 2 Typen von Fehlern unterschieden:

Der Typ I Fehler beziffert die Wahrscheinlichkeit, rein zufällig einen Effekt zu entdecken, obwohl keiner vorhanden ist. Er wird durch den *p-Wert* angegeben, wobei *p* für „Wahrscheinlichkeit", engl. *probability* steht. Nach Übereinkunft ist das Signifikanzniveau beim biostatistischen Gruppenvergleich bei p = 0.05 angesetzt. Ein p = 0.05 bedeutet daher, dass man in einer von 20 gleichartigen Studien einen durchschnittlichen Effekt gemessen hat, obwohl in Wahrheit keiner vorhanden war. Je größer die Vergleichsgruppen sind, desto unwahrscheinlicher treten diese Zufallsschwankungen auf, respektive desto kleiner wird der p-Wert [17]. Die Tatsache, dass eine Studie statistisch signifikante Unterschiede (niedrige p-Werte) vorweist, bedeutet jedoch nicht, dass diese Unterschiede auch von klinischer Relevanz sind. Daher ist es immer sinnvoll darauf zu achten, ob der durch die Intervention erzielte Effekt (die Effektgröße) auch klinische Konsequenzen hat.

Der Typ II Fehler beziffert die Wahrscheinlichkeit, einen Effekt zu übersehen, obwohl er vorhanden war. In diesem Zusammenhang wird die *Power* einer Studie angegeben. Je größer die Power, desto kleiner ist der Typ II Fehler (Power = 1 – Typ II Fehler). Nach Übereinkunft sollte die Power einer Studie nicht unter 80 % liegen, bei dieser Minimalanforderung nimmt man aber in Kauf, in 2 von 10 Fällen einen tatsächlich bestehenden Effekt zu übersehen.

Power, p-Wert, die Gruppengröße sowie die Effektgröße hängen miteinander zusammen. Da der p-Wert einerseits durch das Signifikanzniveau festgelegt ist, und sich andererseits die Effektgröße aus der klinischen Fragestellung heraus ergibt, kann man vor Studienbeginn berechnen, wie groß die Studiengruppen sein müssen, um eine ausreichende Studienpower zu erhalten. Diese Berechnung der Stichprobengröße sollte immer vor Studienbeginn durchgeführt werden, um das Ausmaß der Zufallsvariabilität einzuschränken.

Studiendauer

Die wichtigsten Parameter, an denen medizinische Interventionen gemessen werden, sind Wirksamkeit und Sicherheit. Dies kann nachhaltig jedoch nur nach Beobachtung einer Intervention über einen langen Zeitraum erhoben werden. Da RCTs aber meist nur über einen relativ kurzen Zeitraum durchgeführt werden, können oft nur kurzfristige und relativ häufige unerwünschte Ereignisse dokumentiert werden. Wie man etwa am Beispiel der Hormonersatztherapie sehen kann, lassen sich wirklich verlässliche Aussagen über die Nebenwirkungen einer Therapie aber letztlich nur durch groß angelegte Langzeitstudien erreichen [18].

Externe Validität

Die mangelnde Generalisierbarkeit des Studienergebnisses ist oft einer der Hauptkritikpunkte an RCTs. Wichtigstes Augenmerk ist hierbei auf die Auswahl der Studienpatienten zu legen. Nur sehr wenige Patienten, welche an einer spezifischen Erkrankung leiden, werden in eine entsprechende Studie eingeschlossen. Die in Studien eingeschlossenen Patienten entsprechen häufig nicht jenen Patienten, welche dann die Therapie in der täglichen Praxis erhalten. In vielen RCTs werden beispielsweise ältere Menschen oder Patienten mit häufig vorkommenden Begleiterkrankungen ausgeschlossen [16].

CONSORT Statement

CONSORT ist ein Akronym und steht für Consolidated Standards of Reporting Randomised Controlled Trials [19]. Das CONSORT Statement wurde von klinischen Epidemiologen und Editoren wissenschaftlicher Journale verfasst, um die Qualität von veröffentlichten RCTs zu verbessern und zu vereinheitlichen. Es dient Wissenschaftern als Anleitung für Planung und Präsentation eines RCTs. Die wichtigsten Teile des Statements bestehen aus (a) einer *Checklist* für all jene Informationen, welche in Publikation enthalten sein sollen, und (b) einem *Flow chart*, welcher die Patientenflüsse veranschaulicht, damit Selektions-Bias und externe Validität besser beurteilt werden können (siehe Kapitel 4). In zunehmendem Maße wird von medizinischen Fachjournalen verlangt, dass RCTs gemäß der CONSORT Richtlinien durchgeführt werden. Bei Einreichung wird hierfür eine ausgefüllte Checklist und Flow chart verlangt. Es konnte gezeigt werden, dass sich die Qualität der Publikationen in jenen Journalen deutlich verbesserte, welche CONSORT einführten (British Medical Journal [BMJ], Journal of the American Medical Association [JAMA], The Lancet) nicht aber in anderen Journalen, welche CONSORT nicht einführten (z.B. New England Journal of Medicine[NEJM]) [20]. Mittlerweile wird auch von Autoren, welche bei NEJM einreichen wollen, eine CONSORT-Checklist und -Flowchart verlangt.

Literatur

1. Müllner M (Springer-Verlag Wien, 2005) Erfolgreich wissenschaftlich Arbeiten in der Klinik, Evidence Based Medicine.
2. Last J (2001) A dictionary of epidemiology. New York: Oxford University Press

3. Akobeng AK (2005) Understanding randomised controlled trials. Arch Dis Child 90(8): 840–844
4. Rychetnik L, Hawe P, Waters E, Barratt A, Frommer M (2004) A glossary for evidence based public health. J Epidemiol Community Health 58(7): 538–545
5. Khan K, Riet G, Popay J (2001) Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews; CRD Report 4, Stage II, Phase 5: Study quality assessment. Centre for Reviews and Dissemination
6. Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342(25): 1887–1892
7. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J (2001) Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 286(7): 821–830
8. Day SJ, Altman DG (2000) Statistics notes: blinding in clinical trials and other studies. Bmj 321(7259): 504
9. Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273(5): 408–412
10. Schulz KF, Grimes DA, Altman DG, Hayes RJ (1996) Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. BMJ 312(7033): 742–744
11. Lachin JM (2000) Statistical considerations in the intent-to-treat principle. Control Clin Trials 21(3): 167–189
12. Montori VM, Guyatt GH (2001) Intention-to-treat principle. CMAJ 165(10): 1339–1341
13. Juni P, Altman DG, Egger M (2001) Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 323(7303): 42–46
14. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS (2006) A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 59(10): 1040–1048
15. Green LW, Glasgow RE (2006) Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof 29(1): 126–153
16. Rothwell PM (2005) External validity of randomised controlled trials: „to whom do the results of this trial apply?" Lancet 365(9453): 82–93
17. Moore RA, Gavaghan D, Tramer MR, Collins SL, McQuay HJ (1998) Size is everything—large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. Pain 78(3): 209–216
18. Barrett-Connor E, Grady D, Stefanick ML (2005) The rise and fall of menopausal hormone therapy. Annu Rev Public Health 26: 115–140
19. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134(8): 663–694
20. Moher D, Jones A, Lepage L (2001) Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 285(15): 1992–1995
21. Glossar zur evidenzbasierten Medizin. Deutsches Netzwerk für evidenzbasierte Medizin eV; www.ebm-netzwerk.de

Medieninhaber und Herausgeber: Springer-Verlag GmbH, Sachsenplatz 4–6, 1201 Wien, Österreich – Datenkonvertierung und Umbruch: Manz Crossmedia, 1051 Wien, Österreich, Druck: Druckerei Ferdinand Berger & Söhne Gesellschaft m. b. H., 3580 Horn, Österreich – Verlagsort: Wien. – Herstellungsort: Horn.

Printed in Austria
P. b. b. / Erscheinungsort: Wien / Verlagspostamt 1201 Wien