ÜBERSICHTSARBEIT

Fallzahlplanung in klinischen Studien

Teil 13 der Serie zur Bewertung wissenschaftlicher Publikationen

Bernd Röhrig, Jean-Baptist du Prel, Daniel Wachtlin, Robert Kwiecien, Maria Blettner

ZUSAMMENFASSUNG

Hintergrund: Dieser Artikel beschreibt Ziel, Notwendigkeit und Methodik der Fallzahlplanung in klinischen Studien. Weder zu kleine noch zu große Fallzahlen sind klinisch, methodisch oder ethisch zu rechtfertigen. Die an klinischen Studien beteiligten Mediziner wirken direkt an der Fallzahlplanung mit, da ihre Expertise sowie die Kenntnis der Literatur hierbei unerlässlich sind.

Methode: Anhand einer Auswahl selektiv recherchierter internationaler wissenschaftlicher Artikel und eigener Expertise wird das Vorgehen bei der Fallzahlplanung erläutert.

Ergebnisse: An einem fiktiven Beispiel, in dem unter Verwendung eines t-Tests zwei blutdrucksenkende Medikamente A und B miteinander verglichen werden, wird die Fallzahlplanung dargestellt und beispielhaft berechnet. Anschließend wird ein allgemeines Prinzip zur Fallzahlplanung beschrieben, das grundlegend auch auf andere statistische Tests anwendbar ist. Exemplarisch wird für verschiedene Fälle aufgelistet, welche medizinischen Fachkenntnisse und Annahmen bei der Fallzahlplanung benötigt werden. Diese hängen in der Regel vom statistischen Test ab.

Schlussfolgerung: Jede klinische Studie erfordert eine rationale Begründung für die geplante Stichprobengröße. Eine Fallzahlplanung hat das Ziel, die optimale Probanden- beziehungsweise Patientenzahl für eine klinische Studie zu ermitteln. Geplante Fallzahlen sollten in Zusammenarbeit mit erfahrenen Biometrikern und Medizinern erarbeitet werden. Das medizinische Fachwissen ist aber für die Fallzahlplanung essenziell.

Zitierweise: Dtsch Arztebl Int 2010; 107(31–32): 552–6
DOI: 10.3238/arztebl.2010.0552

Medizinischer Dienst der Krankenversicherung Rheinland-Pfalz (MDK), Referat Rehabilitation/Biometrie: Dr. rer. nat. Röhrig

Institut für Epidemiologie, Universität Ulm: Dr. med. du Prel, MPH

Interdisziplinäres Zentrum Klinische Studien (IZKS), Universitätsmedizin der Johannes Gutenberg Universität Mainz: Dipl.-Kfm. Wachtlin

Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes Gutenberg Universität Mainz: Dr. rer. nat. Kwiecien, Prof. Dr. rer. nat. Blettner

Das Design ist essenziell für die Qualität einer jeden klinischen und epidemiologischen Studie. Die Fallzahlplanung ist dabei ein entscheidender Teil (1). Es ist aus methodischen Gründen notwendig, vor der Durchführung den Ablauf der Studie und die Fallzahl zu bestimmen, und diese vor Beginn der Rekrutierung in einem Protokoll festzulegen. Abweichungen davon sind nur im Rahmen allgemeiner Richtlinien für klinische Studien zulässig. Wird es versäumt, die Fallzahl anzugeben, kann ein unabhängiger Prüfer im Nachhinein nicht mehr feststellen, ob der Experimentator Daten oder statistische Methoden so selektiert hat, dass ein von ihm gewünschtes Resultat „nachgewiesen“ werden konnte. Zudem ist es notwendig, die Wahrscheinlichkeit zu kontrollieren, mit der ein tatsächlich vorhandener Effekt in der Studie als statistisch signifikant entdeckt werden kann. Beispielsweise wird ein pharmazeutisches Unternehmen zur geplanten Einführung eines neuen Medikaments sowohl aus ökonomischen als auch aus ethischen Gründen nicht riskieren, den Nachweis der Wirksamkeit oder der Nichtunterlegenheit gegenüber anderen Medikamenten wegen einer zu geringen Fallzahl nicht erbringen zu können. Ebenso kann es nicht toleriert werden, dass an zu vielen Patienten das neue Medikament untersucht wird. Sowohl Studien mit zu kleiner als auch solche mit zu großer Fallzahl sind somit ethisch und ökonomisch nicht zu rechtfertigen (2–4). Auch bei deskriptiven und retrospektiven Studien sollte vorher geplant werden, aus welchen Quellen und in welchem Umfang Daten gesammelt werden. Die Fallzahlplanung ist in der medizinischen Forschung unumgänglich. Fehlt diese, so spricht das für einen Mangel an Qualität der entsprechenden Studie und die Resultate werden mit Skepsis betrachtet.

Der vorliegende Artikel beschäftigt sich maßgeblich mit der Fallzahlplanung bei vorgesehener Anwendung eines einzelnen statistischen Tests in Bezug auf eine konfirmatorische Fragestellung. Das Ziel der Fallzahlplanung ist es, die Stichprobenumfänge so zu wählen, dass ein tatsächlich vorhandener Effekt mit einer hohen Wahrscheinlichkeit als statistisch signifikant erfasst wird. Zusätzlich geht es darum, genügend Sicherheit zu haben, dass ein solcher Effekt auch tatsächlich nicht existiert, wenn er in der Studie nicht gefunden werden kann (4).

Bestimmung von Fallzahlen

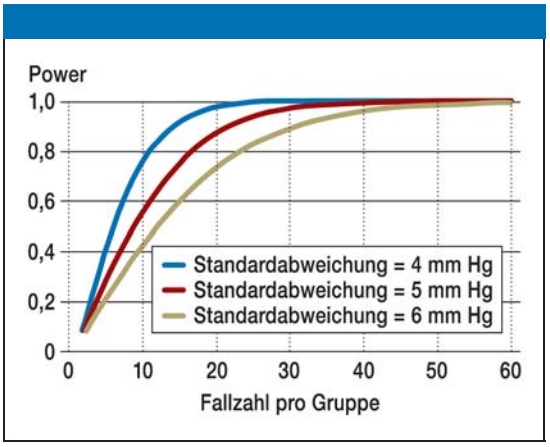
Für eine Studie zum Vergleich von zwei blutdrucksenkenden Medikamenten A und B werden durch Randomisierung der Studienteilnehmer – also zufälliger Zuweisung der Patienten in die Therapiegruppen – zwei homogene

und unabhängige Gruppen gebildet. Die Patienten der ersten Gruppe erhalten Medikament A, die der zweiten Gruppe erhalten Medikament B. Die mittlere Senkung des Blutdrucks nach vier Wochen sei der primäre Endpunkt.

Aus Literaturstudien weiß man, dass die Senkung des Blutdrucks in der Population der Hypertoniker unter beiden Medikamenten als normalverteilt angenommen werden kann und dass das Medikament A den Blutdruck von Hypertonikern im Mittel um etwa 10 mm Hg senkt. Aufgrund bisheriger Untersuchungen wird bei Medikament B eine stärkere Senkung von etwa 15 mm Hg erwartet. Dies wird als eine relevante Verbesserung angesehen. Darüber hinaus wird für beide Medikamente aufgrund medizinischer Einschätzung eine Standardabweichung von 5 mm Hg bei der Blutdrucksenkung unterstellt.

Zur Klärung der Frage, ob das Medikament B den Blutdruck statistisch signifikant stärker senkt als Medikament A, kann ein 1-seitiger t-Test nach Student für unabhängige Stichproben durchführt werden (5, 6). Um weder zu wenige, noch zu viele Patienten in die Studie einzuschließen, wird eine Fallzahlplanung durchgeführt. Zur Bestimmung einer Fallzahl werden die Power (zu deutsch: Teststärke, Macht beziehungsweise Güte) und das Signifikanzniveau (7) des statistischen Tests vorher festgelegt. Für das Signifikanzniveau – das ist die Wahrscheinlichkeit, ein statistisch signifikantes Testergebnis zu erhalten, auch wenn in Wirklichkeit kein Unterschied besteht – ist bei 1-seitigen Tests ein Wert von 2,5 % üblich (vergleiche [8], Sektion 5.5). Je nach Fragestellung sind jedoch auch andere Werte denkbar. Für die Power – das ist die Wahrscheinlichkeit, den tatsächlich vorhandenen Unterschied mit dem statistischen Test aufzudecken – wird oftmals ein Wert von 80 % oder 90 % verwendet.

Die *Grafik* veranschaulicht diese Relation für eine Standardabweichung von 4, 5 und 6 mm Hg. Für eine Standardabweichung von 5 mm Hg ist bei den oben genannten Zahlen und der festgelegten Power von 80 % eine Fallzahl von 17 für jede Gruppe notwendig. Bei einer Standardabweichung von 4 mm Hg wäre eine Fallzahl von 12 Probanden pro Gruppe, bei 6 mm Hg eine Fallzahl von 24 Probanden pro Gruppe nötig *(Grafik)*. Hierzu wird zusätzlich eine kleine Beispielrechnung im *Kasten* vorgestellt.



Power (Teststärke) eines einseitigen t-Tests (Niveau = 2,5 %) in Abhängigkeit – etwa für den Vergleich zweier Medikamente A und B – von der Fallzahl (t-Test für gleiche Standardabweichungen in den beiden Studiengruppen A und B zum Vergleich von Mittelwerten)

Notwendiges medizinisches Fachwissen

Im obigen Beispiel ist zur Schätzung des erwarteten Unterschieds und der Streuung zur blutdrucksenkenden Wirkung beider Medikamente medizinisches Fachwissen notwendig. Zu diesem Zweck dienen häufig Literaturrecherchen oder auch Pilotstudien. Der Biometriker kann dem Mediziner behilflich sein, diese Angaben zu ermitteln. Die inhaltliche Bedeutung kann jedoch nur vom fachkundigen Mediziner beurteilt werden. So ist es die Aufgabe des Mediziners, und nicht des Biometrikers, zu entscheiden, ob der erwartete Unterschied bezüglich der mittleren Blutdrucksenkung beider Medikamente auch klinisch bedeutsam ist. Unterscheiden sich die Medikamente beispielsweise nur um 1 mm Hg, könnte daraus vermutlich nicht abgeleitet werden, dass die Patienten, die mit dem stärker blutdrucksenkenden Präparat behandelt werden, von dieser Behandlung auch einen Vorteil, zum Beispiel im Sinne eines verringerten Risikos kardiovaskulärer Ereignisse, haben.

Die vorgestellte Vorgehensweise zur Bestimmung von Fallzahlen ist auch prinzipiell für andere Tests wie zum Beispiel den unverbundenen Wilcoxon-Rangsummen-Test auf Lageunterschied oder den exakten Fisher-Test zum Vergleich zweier Raten möglich. Je nach statistischem Verfahren werden unterschiedliche Informationen vom Mediziner benötigt. In *Tabelle 1* wird exemplarisch für einige statistische Verfahren aufgelistet, welche Annahmen eine Fallzahlplanung ermöglichen.

Beim t-Test sollte der Mediziner Annahmen über die Mittelwerte (μ_1 und μ_2) in zwei Populationen sowie Annahmen über die Standardabweichungen (σ_1 und σ_2) in diesen Populationen liefern.

Für den Fisher-Test sind Schätzungen über die relativen Anteile beziehungsweise die Raten von Ereignissen (π_1 und π_2) in beiden Populationen ausreichend. Dazu muss aus der Literatur ermittelt werden, bei wie vielen von jeweils 100 Patienten unter Therapie 1 und Therapie 2 in etwa ein Ereignis, wie beispielsweise Nebenwirkungen, auftritt (= relative Häufigkeiten).

Für den Wilcoxon-Rangsummentest ist eine fachkundige Schätzung zur Wahrscheinlichkeit, dass die Zielvariable zur zufälligen Ziehung aus Population 1 kleiner ist als die zufällige Zielvariable aus Population 2, nötig. Eine Schätzung beziehungsweise Annahme für diese Größe sollte unbedingt in Zusammenarbeit mit einem Biometriker erstellt werden.

Eine sorgsame Einschätzung der notwendigen Parameter ist lohnend und kann fehlerhaften Poweranalysen und Fallzahlberechnungen erheblich vorbeugen (9).

Fallzahlplanung

Das genannte Beispiel zum unverbundenen t-Test veranschaulicht ein häufig verwendetes Schema zur Bestimmung von Fallzahlen. Nach einer Einschätzung notwendiger Parameter, zum Beispiel Mittelwerte und Standardabweichungen, und der Festlegung eines Signifikanzniveaus, werden für variierende Annahmen zur Power die Fallzahlen zum entsprechenden Test ermittelt. Dabei handelt es sich um folgende Relation: Je größer die

KASTEN

Beispielrechnung

Für den einseitigen, unverbundenen t-Test gilt zur Vereinfachung die Restriktion gleich großer Gruppengrößen $n_1 = n_2$ und gleicher Standardabweichungen $\sigma_1 = \sigma_2 = \sigma$ in den beiden Populationen, die auf Mittelwertunterschiede untersucht werden sollen. Die Mittelwertunterschiede zwischen den beiden Populationen werden als $\mu_1 - \mu_2$ bezeichnet. Für die gewünschte „Power" wird gewöhnlich 0,8 = 80 % gesetzt, α gibt das Signifikanzniveau wieder. Hier soll $n = n_1 + n_2$ gesetzt werden. Ziel ist es, die Gesamtstichprobe N zu bestimmen. Für die Fallzahlplanung kann folgende, vereinfachte, approximative Faustformel verwendet werden (allerdings geht mit der approximativen Vereinfachung eine gewisse Ungenauigkeit einher):

$$n \approx \left[\frac{2(z_{Power} + z_{1-\alpha})}{2(\mu_1 - \mu_2)/\delta}\right]^2$$

Hier bezeichnet $z_{1-\alpha}$ das 1-α-Quantil der Standardnormalverteilung, dessen Werte aus statistischen Tabellen entnommen werden können. Um den Stichprobenumfang für den unverbundenen t-Test zu ermitteln, wird in dieser Relation einfach α durch α/2 ersetzt, ansonsten ist die Vorgehensweise gleich. Diese Relation ist in (16, Kapitel 12) zu finden.

Beispiel:
Hier soll obige Relation zur Bestimmung der Fallzahl aus dem Beispiel der blutdrucksenkenden Medikamente A und B mit einer angenommenen erwarteten Mittelwertdifferenz von 5 mm Hg und einer angenommenen gemeinsamen Standardabweichung von 6 mm Hg bei Einnahme von Medikament A oder B verwendet werden. Es sollen der einseitige, unverbundene t-Test zum Signifikanzniveau von 2,5 % eingesetzt und die Fallzahl so bestimmt werden, dass dieser Test eine „Power" von 80 % hat. Aus statistischen Tabellen ist abzulesen, dass $z_{0,8} = 0{,}8416$ und $z_{0,975} = 1{,}96$ (vergleiche zum Beispiel [17]). Diese Zahlen, in obige Relation eingesetzt, ergeben für den gesamten Stichprobenumfang:

$$45{,}2 \approx \left[\frac{2(0{,}8416 + 1{,}96)}{5/6}\right]^2$$

Die Stichproben sollen nach Voraussetzung obiger Formel gleich groß sein. Die einzelnen Stichproben sollen also etwa einen Umfang von 22,6 = 45,2/2 haben. Demnach werden 23 Probanden in jeder Gruppe benötigt. Eine exaktere Kalkulation ergibt allerdings 24 Probanden pro Gruppe.

TABELLE 1

Notwendige Annahmen zur Fallzahlplanung oder Poweranalyse bei verschiedenen Tests zum Vergleich von zwei Populationen

Testverfahren	Medizinische Annahme
unverbundener t-Test bei verschiedenen Standardabweichungen	Standardabweichungen σ_1, σ_2 Mittelwerte μ_1, μ_2
unverbundener Wilcoxon-Rangsummentest	Wahrscheinlichkeit P ($X_1 < X_2$)
exakter Fisher-Test zum Vergleich zweier Raten	relative Anteile π_1, π_2

Power – also die Sicherheit, ein signifikantes Ergebnis zu erhalten –, desto größer ist die benötigte Fallzahl der Studie. Gewählt wird also die kleinste Fallzahl, so dass eine vorgegebene Power mindestens erreicht wird.

Andererseits kommt es auch vor, dass eine Fallzahl durch externe Faktoren beschränkt ist – beispielsweise durch die Dauer der Rekrutierungszeit, seltene Krankheiten oder die zeitliche Limitierung einer geförderten Studie –, und dennoch die Auswertung durch einen statistischen Test vorgesehen ist. In diesem Fall muss bei der Planung die erreichbare Power ermittelt werden. Je geringer die Power ist, desto aussichtsloser ist es, die vermutete Hypothese nachzuweisen (2, 3). Eine zu geringe Power kann dazu führen, dass eine Studie in der Planung modifiziert oder nicht durchgeführt wird. Breckenkamp und Koautoren (10) berichten von einer geplanten Kohortenstudie, in der der Zusammenhang zwischen der beruflichen Exposition gegenüber elektromagnetischen Feldern und Krebserkrankungen untersucht werden sollte. Die Autoren geben an, dass in keiner der denkbaren beruflichen Kohorten genügend Personen exponiert gewesen wären. So wurde keine Studie durchgeführt, obwohl eine solche Untersuchung aus umweltpolitischer Sicht erwünscht war.

Steht nicht der Nachweis einer Hypothese im Mittelpunkt einer Studie, sondern die Schätzung eines Parameters, dann kann eine Strategie der Fallzahlplanung zur Abschätzung der erwarteten Breite von Konfidenzintervallen verfolgt werden (7). Angenommen, es soll die Prävalenz der Personen mit erhöhtem Blutdruck geschätzt werden (zuzüglich eines 95-%-Konfidenzintervalls). Je kleiner das Konfidenzintervall ist, desto besser kann dieser Populationsparameter (hier Prävalenz) eingegrenzt werden. Durch die Festlegung der erwarteten Breite dieses Konfidenzintervalls kann eine Fallzahl bestimmt werden. Bei einem solchen Verfahren ist es notwendig, eine Vorstellung der Größe der Prävalenz und eine gewünschte Präzision vorzugeben.

Da auch mit medizinischem Fachwissen häufig nur grobe, recht unzuverlässige Einschätzungen der in die Bestimmung von Fallzahlen eingehenden Parameter möglich sind, werden oftmals mehrere Szenarien untersucht. Dazu sollen exemplarisch nochmals das genannte Beispiel und die *Grafik* betrachtet werden. Bei einer angenommenen Standardabweichung von 5 mm Hg waren für eine Power von 80 % insgesamt 17 Probanden pro Gruppe notwendig. Liegt die Standardabweichung wider Erwarten bei 6 mm Hg, dann beträgt die Power nur noch 65 % und erst bei einer Erhöhung auf 24 Probanden pro Gruppe wieder etwa 80 %. Hier ist zu sehen, dass eine Erhöhung der Streuung auch eine Erhöhung der Fallzahl zur Folge hat. Auch eine Verringerung des Signifikanzniveaus führt zu höheren Fallzahlen, weil dadurch die Fehlerwahrscheinlichkeit, den Effekt irrtümlicherweise nachzuweisen, kleiner wird. Das Signifikanzniveau darf allerdings nicht zum Zweck der Fallzahlplanung variiert werden. Weitere Relationen dieser Art sind in *Tabelle 2* anhand des unverbundenen t-Tests veranschaulicht.

Darüber hinaus sollte stets beachtet werden, dass ein nachzuweisender Unterschied auch klinisch relevant ist.

Die um 5 mm Hg deutlichere Senkung bei Medikament B im Vergleich zu Medikament A wird vom Kliniker/Forscher als klinisch relevanter Effekt angesehen. Ist der in der Studie zu erwartende Effekt aber zu klein, dann ist der Nutzen der klinischen Studie in Frage zu stellen. In diesem Fall könnten auch statistisch signifikante Ergebnisse klinisch nicht relevant sein (7).

Ein wesentlicher Punkt bei der Fallzahlplanung ist die Berücksichtigung von „Lost-to-Follow-up“ beziehungsweise „Drop-out“ (11). Ist beispielsweise davon auszugehen, dass bei einem Teil der Probanden in einer Studie – aus welchen Gründen auch immer – keine hinreichende Datenerhebung möglich sein wird, so muss die Fallzahl entsprechend diesem Anteil erhöht werden. Um wie viele Patienten die Fallzahl angehoben werden muss, hängt von der geschätzten Teilnehmerrate und den Studienbedingungen ab. Es sei allerdings darauf hingewiesen, dass solche Gegebenheiten in der Regel auch die Repräsentativität der Daten beeinflussen. Eine Verzerrung der Resultate ist in der Regel die Folge. Dies ist bei der Planung der Studie ebenfalls zu berücksichtigen.

Für die häufigsten Tests stehen explizite Formeln zur Bestimmung von Fallzahlen zur Verfügung (12–14). Machin und Koautoren (12) liefern für übliche Werte von Größen, die in die Fallzahlplanung eingehen, umfangreiche Tabellenwerke, aus denen die Fallzahl direkt abgelesen werden kann.

Als gängige Statistiksoftwareprogramme liefern SPSS mit SamplePower und SAS mit den Prozeduren „PROC POWER“ und „PROC GLMPOWER“ sowie die Software Nquery für die Berechnung von Fallzahlen geeignete Lösungen. Kostenlos kann das Programm G*Power 3 des Instituts für experimentelle Psychologie der Heinrich Heine Universität Düsseldorf benutzt werden (www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/). Es empfiehlt sich, ein validiertes Programm – wie zum Beispiel eines der oben genannten – zu verwenden.

TABELLE 2

Auswirkungen von Veränderungen verschiedener Größen auf die Fallzahl anhand des 1-seitigen unverbundenen t-Tests nach Student unter Annahme gleicher Standardabweichungen

Veränderung	Effekt*1	Standard-abweichung	Effekt-stärke*2	Signifikanz-niveau	Power	Fallzahl (pro Gruppe)
Effekt	5	5	1,0	0,025	0,80	17
	3	5	0,6	0,025	0,80	46
	1	5	0,2	0,025	0,80	401
	0,5	5	0,1	0,025	0,80	1 600
Standardabweichung	5	25	0,2	0,025	0,80	401
	5	10	0,5	0,025	0,80	65
	5	8	0,625	0,025	0,80	42
	5	3	1,666	0,025	0,80	7
Signifikanzniveau	5	5	1,0	0,05	0,80	14
	5	5	1,0	0,01	0,80	22
	5	5	1,0	0,001	0,80	34
Power	5	5	1,0	0,025	0,95	27
	5	5	1,0	0,025	0,90	23
	5	5	1,0	0,025	0,70	14

*1 Effekt: Differenz der beiden Miittelwerte; *2 Effektstärke: Effekt dividiert durch die Standardabweichung

Diskussion

Zur Planung der Fallzahl einer klinischen Studie braucht man Vorinformationen. Welche Vorinformationen notwendig sind, hängt von den geplanten statistischen Methoden ab. Können die entsprechenden Größen nicht geschätzt werden, empfiehlt es sich beispielsweise, vor der konfirmatorischen Studie eine Pilotstudie durchzuführen, mit dem Ziel, die entsprechenden Parameter der Populationen zu schätzen. Auf alle Fälle sollte der erwartete Effekt mindestens so groß wie der minimale klinisch relevante Effekt sein.

Auch bei explorativen und deskriptiven Studien (1) muss der Umfang der Studiengruppe(n) bestimmt werden, um die zu schätzenden Parameter ausreichend genau eingrenzen zu können. Eine fehlende Fallzahlplanung spricht für eine schlechte Qualität einer Studie.

Die Fallzahlplanung für eine klinische Studie basiert auf einer Abschätzung aufgrund von Vorinformationen, die von Studie zu Studie auch unterschiedlich präzise sein kann. Dies sollte bei der Interpretation der Ergebnisse stets berücksichtigt werden. Ein in der Planungsphase überschätzter Behandlungseffekt hat in der Regel eine zu geringe Fallzahl zur Folge. Der beobachtete Behandlungseffekt kann dann lediglich wegen der zu geringen Fallzahl nicht signifikant sein.

Bei jeder Fallzahlplanung sollten auch der Umgang mit fehlenden Werten und aus der Studie ausscheidende Patienten berücksichtigt werden.

Nur ein kleiner Ausschnitt der Fallzahlplanung kann hier beleuchtet werden. Je nach Studiendesign gibt es aber noch weitere Aspekte, die bei der Fallzahlplanung wichtig sind. Die Methoden der Fallzahlplanung können sich beispielsweise ändern, wenn bei der klinischen Studie ein Test auf Überlegenheit, auf Nicht-Unterlegenheit oder auf Äquivalenz durchgeführt werden soll (13). Bei Nicht-Unterlegenheitsstudien können recht hohe Fallzahlen erforderlich sein, da als nachzuweisender mittlerer Unterschied oftmals der kleinste klinisch relevante Unterschied angesetzt wird, der dann als Nicht-Unterlegenheitsschranke fungiert. Dieser ist in der Regel wesentlich kleiner, als ein tatsächlicher mittlerer Unterschied.

Oftmals sollen anhand eines Datensatzes mehrere Hypothesen geprüft werden. Multiple Testprobleme müssen bei der Fallzahlplanung berücksichtigt werden. Vielfach wird daher nur eine Hauptfragestellung festgelegt.

Darüber hinaus ist die Fallzahl bei modernen Studien nicht immer determiniert. Beispielsweise kann im Rahmen adaptiver Designs nach einem in der Planungsphase streng festgelegten Schema die Fallzahl während einer Studie beeinflusst oder gesteuert werden. Dieses Vorgehen erfordert aber eine sorgfältige, statistisch anspruchsvolle Planung und sollte nie ohne einen erfahrenen Biometriker durchgeführt werden.

Aufgrund der Komplexität und der weitreichenden Folgen der Fallzahlplanung empfiehlt sich die Zusammenarbeit zwischen erfahrenen Biometrikern und Medizinern. Durch die gemeinsame Planung aller wichtigen Details kann die Qualität und Aussagekraft von Studien entscheidend verbessert werden (2, 3, 15).

KERNAUSSAGEN

- Fallzahlplanung ist ein unumgänglicher Schritt bei der Durchführung klinischer Studien.
- Für eine Fallzahlplanung ist die Expertise des Mediziners zur quantitativen Schätzung der relevanten Effekte notwendig.
- Die Fallzahlplanung hängt von der geplanten statistischen Auswertungsmethode und damit von der medizinischen Fragestellung ab.
- Die Erfolgschancen einer klinischen Studie und die Qualität der Forschungsresultate hängen maßgeblich von der Fallzahlplanung ab.
- Die Planung von Fallzahlen sollte stets in Zusammenarbeit mit einem fachkundigen Statistiker beziehungsweise Biometriker erfolgen.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt im Sinne der Richtlinien des International Committee of Medical Journal Editors besteht.

Manuskriptdaten
eingereicht: 15. 1. 2010, revidierte Fassung angenommen: 22. 3. 2010

LITERATUR

1. Röhrig B, du Prel JB, Blettner M: Study design in medical research – Part 2 of a series on evaluation of scientific publications [Studiendesign in der medizinischen Forschung. Teil 2 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 2009; 106(11): 184–9.
2. Eng J: Sample size estimation: how many individuals should be studied? Radiology 2003; 227: 309–13.
3. Halpern SD, Karlawish JHT, Berlin JA: The continuing unethical conduct of underpowered clinical trails. JAMA 2002; 288: 358–62.
4. Altman DG: Practical Statistics for medical research. London: Chapman and Hall 1991.
5. du Prel JB, Röhrig B, Hommel G, Blettner M: Choosing Statistical Tests. Part 12 of a series on evaluation of scientific publications [Auswahl statistischer Testverfahren: Teil 12 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 2010; 107(19): 343–8.
6. Sachs L: Angewandte Statistik: Anwendung statistischer Methoden. 11th edition. Springer 2004; 352–61.
7. du Prel JB, Hommel G, Röhrig B, Blettner M: Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications [Konfidenzintervall oder p-Wert? Teil 4 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 2009; 106(19): 335–9.
8. ICH E9: Statistical Principles for Clinical Trials. London UK: International Conference on Harmonization 1998; adopted by CPMP July 1998 (CPMP/ICH/363/96).
9. Blettner M, Ashby D: Power calculation for cohort studies with improved estimation of expected numbers of death. Soz Präventivmed 1992; 37: 13–21.
10. Breckenkamp J, Berg-Beckhoff G, Münster E, Schüz J, Schlehofer B, Wahrendorf J, Blettner M: Feasibility of a cohort study on health risks caused by occupational exposure to radiofrequency electromagnetic fields. Environ Health 2009; 8: 23.
11. Schumacher M, Schulgen G: Methodik klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung (Statistik und Ihre Anwendungen). 3rd edition. Berlin, Heidelberg, New York: Springer Verlag 2008: 1–436.
12. Machin D, Campbell MJ, Fayers PM, Pinol APY: Sample size tables for clinical studies. 2nd edition. Oxford, London, Berlin: Blackwell Science Ltd. 1987; 1–315.
13. Chow SC, Shao J, Wang H: Sample size calculations in clinical research. Boca Raton: Taylor & Francis, 2003; 1–358.
14. Bock J: Bestimmung des Stichprobenumfangs für biologische Experimente und kontrollierte klinische Studien. München: Oldenbourg Verlag 1998; 1–246.
15. Altman DG: Statistics and ethics in medical research, misuse of statistics is unethical, BMJ 1980; 281: 1182–4.
16. Altman DG, Machin D, Bryant TN, Gardner MJ: Statistics with confidence. 2nd edition. BMJ Books 2000.
17. Fahrmeir L, Künstler R, Pigeot I, Tutz G: Statistik: Der Weg zur Datenanalyse. 4th edition. Berlin, Heidelberg, New York: Springer Verlag 2003; 1–608.

Anschrift für die Verfasser
Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI)
Klinikum der Universität Mainz
Obere Zahlbacher Straße 69
55131 Mainz
E-Mail: blettner-sekretariat@imbei.uni-mainz.de

SUMMARY

Sample Size Calculation in Clinical Trials—Part 13 of a Series on Evaluation of Scientific Publications

Background: In this article, we discuss the purpose of sample size calculation in clinical trials, the need for it, and the methods by which it is accomplished. Study samples that are either too small or too large are unacceptable, for clinical, methodological, and ethical reasons. The physicians participating in clinical trials should be directly involved in sample size planning, because their expertise and knowledge of the literature are indispensable.

Methods: We explain the process of sample size calculation on the basis of articles retrieved by a selecitve search of the international literature, as well as our own experience.

Results: We present a fictitious clinical trial in which two antihypertensive agents are to be compared to each other with a t-test and then show how the appropriate size of the study sample should be calculated. Next, we describe the general principles of sample size calculation that apply when any kind of statistical test is to be used. We give further illustrative examples and explain what types of expert medical knowledge and assumptions are needed to calculate the appropriate sample size for each. These generally depend on the particular statistical test that is to be performed.

Conclusion: In any clinical trial, the sample size has to be planned on a justifiable, rational basis. The purpose of sample size calculation is to determine the optimal number of participants (patients) to be included in the trial. Sample size calculation requires the collaboration of experienced biostatisticians and physician-researchers: expert medical knowledge is an essential part of it.

Zitierweise: Dtsch Arztebl Int 2010; 107(31–32): 552–6
DOI: 10.3238/arztebl.2010.0552

The English version of this article is available online:
www.aerzteblatt-international.de