M. Kutschmann[1]
R. Bender[2]
U. Grouven[2]
G. Berg[1]

Aspekte der Fallzahlkalkulation und Powerberechnung anhand von Beispielen aus der rehabilitationswissenschaftlichen Forschung

Aspects of Sample Size Determination and Power Calculation Illustrated on Examples from Rehabilitation Research

Zusammenfassung

Vielfach wird in medizinischen Studien berichtet, dass ein vermuteter Effekt, z. B. bezüglich der Wirksamkeit einer Maßnahme, nicht gefunden werden konnte. Dies kann damit zusammenhängen, dass die Zahl der in die Studie eingeschlossenen Patienten zu klein war, um einen tatsächlich vorhandenen Effekt entdecken zu können. Oft ist dies darauf zurückzuführen, dass vor Beginn der Studie eine solide Fallzahlkalkulation nicht durchgeführt wurde. Damit fehlen Informationen darüber, wie viele Patienten man hätte einschließen müssen, um den vermuteten Effekt, sofern er vorhanden ist, auch nachweisen zu können. Auf der anderen Seite besteht die Gefahr einer fehlenden Fallzahlkalkulation darin, dass mehr Personen als nötig in die Studie eingeschlossen werden. Dies ist aus zeit- und kostenökonomischen, insbesondere aber auch aus ethischen Gründen bedenklich. Im vorliegenden Beitrag wird das Prinzip der Fallzahlkalkulation erläutert und auf seine Bedeutung – insbesondere in der rehabilitationswissenschaftlichen Forschung – eingegangen.

Schlüsselwörter

Power · Fallzahl · Stichprobenumfang · Signifikanz · relevanter Unterschied

Abstract

Often it is reported in medical studies that an expected effect could not be detected. This may be the case if the sample size had been too small to detect an effect which actually exists. This often is due to the fact that sound sample size estimation had been omitted prior to the study outset. As a result, it is not known how many persons should have been involved in the study to detect this effect if present. On the other hand, if sample size estimation has not been realized, more persons than needed might be included in the study. This is problematic for economic and in particular for ethical reasons. The aim of this paper is to point out the principles of sample size estimation as well as to emphasize its importance not only in general but also in medical rehabilitation research.

Key words

Power · sample size · significance · relevant difference

Koordinatoren der Reihe „Methoden in der Rehabilitationsforschung“:
Prof. Dr. Dr. Hermann Faller, Würzburg; Prof. Dr. Thomas Kohlmann, Greifswald;
Dr. Christian Zwingmann, Siegburg
Interessenten, die einen Beitrag zur Reihe beisteuern möchten, werden gebeten, vorab Kontakt aufzunehmen, E-mail: christian.zwingmann@web.de

Institutsangaben
[1] AG Epidemiologie & International Public Health, Fakultät für Gesundheitswissenschaften, Universität Bielefeld
[2] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

Korrespondenzadresse
Dr. Marcus Kutschmann · Universität Bielefeld · Fakultät für Gesundheitswissenschaften · AG Epidemiologie & International Public Health · Universitätsstraße 25 · 33615 Bielefeld
E-mail: marcus.kutschmann@uni-bielefeld.de

Bibliografie
Rehabilitation 2006; 45: 377 – 384 © Georg Thieme Verlag KG Stuttgart · New York
DOI 10.1055/s-2006-940113
ISSN 0034-3536

Einleitung

Jakob Bernoulli formuliert in seiner *Ars Conjectandi:* „Jedem ist klar, dass es zur Beurteilung irgendeiner Erscheinung nicht ausreicht, eine oder zwei Beobachtungen zu machen, sondern es ist eine große Anzahl von Beobachtungen erforderlich" (zit. nach [1]). Allerdings stellt sich die Frage, wie viele Beobachtungen es genau sein müssen, wenn einem daran gelegen ist, besagte „Erscheinung" auf wissenschaftlich angemessene Weise zu beurteilen. Wie viele Patienten müssen also z. B. in einer rehabilitationswissenschaftlichen Studie eingeschlossen werden, wenn man herausfinden möchte, ob ein neu entwickeltes Nachsorgeprogramm nach einer Rehabilitation berufliche Eingliederungsprozesse verbessert gegenüber dem Fehlen dieses Programms oder ob eine effizientere Gestaltung eines Rehabilitationsprogramms zu einer Erhöhung der gesundheitsbezogenen Lebensqualität eher beiträgt als ein herkömmliches? Um diese Fragen adäquat beantworten zu können, sollte *vor* Beginn der Studie unter Beachtung bestimmter statistischer Gesetzmäßigkeiten der erforderliche Stichprobenumfang berechnet werden. Die für das Verständnis der Fallzahlkalkulation notwendigen Vorüberlegungen und Gesetzmäßigkeiten werden im Folgenden dargestellt.

Zentrale Begriffe

Installiert man einen Feuermelder, möchte man vermeiden, dass er Alarm schlägt, obwohl es gar nicht brennt. Kein Richter möchte jemanden verurteilen, der eigentlich unschuldig ist. Und genauso wenig möchte man in der rehabilitationswissenschaftlichen Forschung einen Unterschied – z. B. zwischen zwei Therapieformen – irrtümlicherweise entdecken, der eigentlich gar nicht vorhanden ist. Ist dennoch ein Fehlalarm zu verzeichnen, wird dennoch ein Unschuldiger verurteilt und wird irrtümlich ein Unterschied zwischen zwei Therapieformen nachgewiesen, begeht man – in der Terminologie der Statistik – einen so genannten „Fehler 1. Art".

Genauso wenig wie einen Fehler 1. Art, möchte man einen „Fehler 2. Art" begehen. Dieser liegt dann vor, wenn der Feuermelder *keinen* Alarm schlägt, *obwohl* es brennt, oder der Richter einen Schuldigen freispricht. Man begeht einen Fehler 2. Art auch dann, wenn in der rehabilitationswissenschaftlichen Forschung ein tatsächlich vorhandener relevanter Unterschied zwischen zwei Therapieformen übersehen wird.

Statt des Fehlers 2. Art kann auch das entsprechende „Gegenereignis" betrachtet werden, das in dem Maße, in dem der Fehler 2. Art vermieden werden sollte, wünschenswert ist. So sollte man nach Möglichkeit einen Feuermelder installieren, der dann Alarm schlägt, wenn es tatsächlich brennt. Ebenso wünschenswert wäre es, wenn in der Judikative Richter tätig sind, die Angeklagte dann verurteilen, wenn sie ein Verbrechen begangen haben. Und beim Vergleich zweier Therapieformen möchte man einen signifikanten Unterschied dann entdecken, wenn er auch tatsächlich vorhanden ist (Feuermelder- und Richteranalogien aus [2] und [3]).

Es liegt auf der Hand, dass man sowohl den Fehler 1. Art als auch den Fehler 2. Art nach Möglichkeit vermeiden sollte oder – anders formuliert – die *Wahrscheinlichkeiten* dafür, dass man diese Fehler begeht, möglichst klein hält. In der Statistik wird die Wahrscheinlichkeit für den Fehler 1. Art durch Angabe des sog. Signifikanzniveaus α nach oben begrenzt. Zur Festlegung des Signifikanzniveaus wird üblicherweise einer der Werte 0,05, 0,025 oder auch 0,01 verwendet (siehe dazu z. B. auch [4]). Das Signifikanzniveau $\alpha = 0{,}025$ ist hierbei insbesondere bei einseitiger Testformulierung gebräuchlich (siehe unten). Je gravierender die Folgen eines Fehlers 1. Art sind, desto kleiner sollte das Signifikanzniveau festgelegt werden. Der am häufigsten verwendete Wert ist allerdings $\alpha = 0{,}05$.

Hält man die Wahrscheinlichkeit β für das Eintreten eines Fehlers 2. Art klein, führt dies zwangsläufig dazu, dass die Wahrscheinlichkeit für das Eintreten des Gegenereignisses $1 - \beta$ groß wird. Diese Wahrscheinlichkeit wird als „Power" bezeichnet. Gebräuchliche Werte für β sind 0,10 bzw. 0,20, sodass $1 - \beta$ den Wert 0,80 bzw. 0,90 annimmt. Die Wahrscheinlichkeit, einen Unterschied zwischen den beiden Therapieformen zu entdecken, *wenn er tatsächlich vorhanden ist,* beträgt also 80 % bzw. 90 %.

Abb. **1** verdeutlicht den Zusammenhang zwischen Fehler 1. Art, Fehler 2. Art (die auch als Alpha- und Betafehler bezeichnet werden) und den entsprechenden Gegenereignissen.

Da das Signifikanzniveau vor Durchführung eines statistischen Tests vorgegeben wird, lässt sich die Power nur noch über den Stichprobenumfang beeinflussen.[1] Und ebenso, wie ein Feuermelder einen Brand nur dann erkennt, wenn umfangreiche Justierungen vorgenommen wurden, oder ein Richter einen Angeklagten nur dann zu Recht verurteilen kann, wenn die Beweisaufnahme lückenlos ist, lässt sich ein vorhandener Unterschied zwischen zwei Therapieformen nur dann nachweisen, wenn ausreichend viele Patienten in die Studie eingeschlossen werden.

Hier stellen sich allerdings Fragen wie: Was sind „ausreichend viele" Patienten? Wie viele Patienten müssen mindestens in die Studie eingeschlossen werden? Ist es sinnvoll, so viele Patienten wie möglich in die Studie einzubeziehen? Letztere Frage muss mit einem klaren „Nein!" beantwortet werden. Dagegen sprechen nicht nur ökonomische Aspekte, da u. U. ein größerer – zeitlicher und/oder finanzieller – Aufwand betrieben wird, als eigentlich notwendig ist. Viel problematischer sind hier ethische Gesichtspunkte. So werden unnötig viele Patienten in der Kontrollgruppe einer möglicherweise unterlegenen Therapie unterzogen. Zu klein darf der Stichprobenumfang allerdings auch nicht sein, da eine zu kleine Fallzahl u. U. dazu führt, dass ein tatsächlich vorhandener Effekt, den eine Therapiemaßnahme möglicherweise hat, schlichtweg übersehen wird. Auch dies ist ethisch bedenklich, da Patienten durch die Teilnahme an einer Studie unnötig belastet werden, obwohl im Grunde von Anfang an klar ist, dass ihr potenzieller Nutzen nicht nachgewiesen werden kann. Des Weiteren hätte eine Therapiemaßnahme, die eigentlich ihren Zweck erfüllt, so keine Chance, sich zu etablieren und zum Wohle der Patienten zu wirken (siehe auch Abschnitt

[1] Der Zusammenhang dieser drei Größen wird im Abschnitt „Zusammenhang zwischen Fallzahl, Power, Signifikanzniveau, relevantem Unterschied und Zielgrößenvariabilität" noch ausführlicher erläutert.

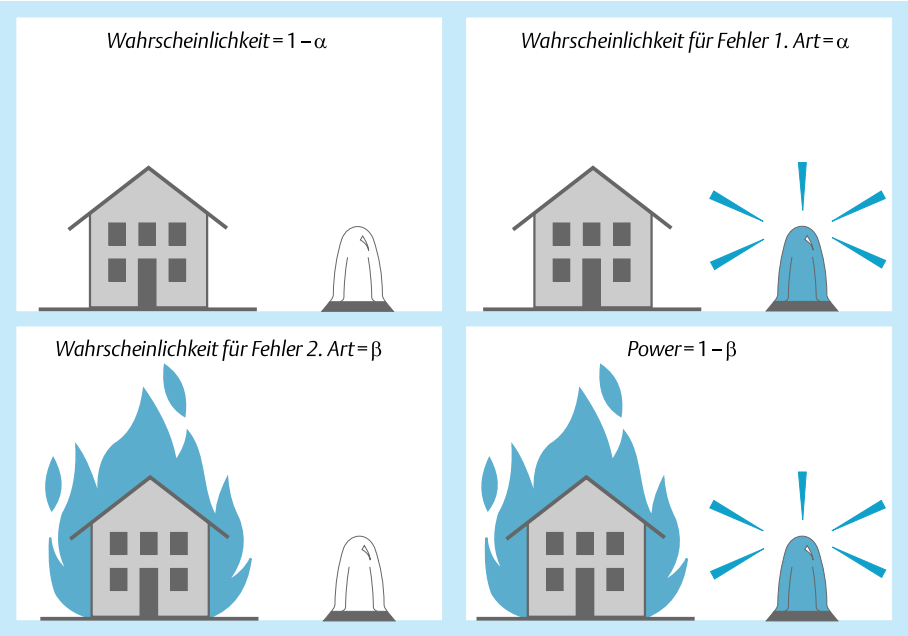

Abb. **1** Zusammenhang zwischen Fehler 1. Art, Fehler 2. Art und Power.

„Konsequenzen einer zu kleinen Fallzahl"). Der Aufwand, der auch bei Studien mit eher kleinem Stichprobenumfang zu betreiben ist, war außerdem vollkommen umsonst, was zumindest aus ökonomischer Sicht bedenklich ist.

Ausgangspunkt jeder Studie sollte eine klar formulierte Fragestellung sein. Eine einfache Formulierung lautet z. B.: „Ist der mittlere Therapieeffekt bei Personen, die einer bestimmten Intervention unterzogen werden, anders als bei Personen, die diese Interventionsmaßnahme nicht erhalten?" Als sog. Nullhypothese H_0 wird dann formuliert: „Es *gibt* im Mittel *keinen* Unterschied zwischen Personen *mit* Intervention und Personen *ohne* Intervention." Die Alternativhypothese H_1 lautet: „Es *gibt* im Mittel einen Unterschied zwischen Personen *mit* Intervention und Personen *ohne* Intervention." Dabei ist darauf zu achten, dass man als Alternativhypothese genau das formuliert, was man eigentlich nachweisen möchte. Formal dargestellt hat dieses statistische Testproblem die Gestalt

Formel 1

$$H_0 : \mu_{mI} = \mu_{oI} \text{ vs. } H_1 : \mu_{mI} \neq \mu_{oI}$$

was äquivalent ist mit der Formulierung

Formel 2

$$H_0 : \mu_{mI} - \mu_{oI} = 0 \text{ vs. } H_1 : \mu_{mI} - \mu_{oI} \neq 0$$

wobei μ_{mI} bzw. μ_{oI} den mittleren Therapieeffekt bei Patienten der Grundgesamtheit mit bzw. ohne Interventionsmaßnahme beschreibt.

Ganz allgemein lässt sich die übliche zweiseitige Hypothesenformulierung statistischer Signifikanztests darstellen durch

Formel 3

$$H_0 : \Theta = \Theta_0 \text{ vs. } H_1 : \Theta \neq \Theta_0$$

wobei Θ ein geeignetes Effektmaß (oben gerade die Differenz der beiden Mittelwerte $\mu_{mI} - \mu_{oI}$) ist und Θ_0 der entsprechende Nullwert (beim Effektmaß $\Theta = \mu_{mI} - \mu_{oI}$ ist $\Theta_0 = 0$). Mit Hilfe von Formel 3 lässt sich erkennen, warum diese Hypothesenformulierung zweiseitig heißt: Die möglichen Werte des Effektmaßes Θ liegen bei der Alternativhypothese auf *beiden* Seiten des Nullwerts Θ_0.

Prinzipiell lässt sich eine statistische Hypothese auch einseitig formulieren durch

Formel 4

$$H_0 : \Theta \leq \Theta_0 \text{ vs. } H_1 : \Theta > \Theta_0$$

In diesem Fall liegen die möglichen Werte des Effektmaßes Θ bei der Alternativhypothese nur auf *einer* Seite des Nullwerts Θ_0. Die Verwendung einseitiger Hypothesen hat bei der Studienplanung durchaus einen Stellenwert, da bei einseitiger Formulierung (bei gleichem Signifikanzniveau) weniger Patienten benötigt werden als bei der zweiseitigen Formulierung [5]. Allerdings ist die Gegenrichtung mit denselben Daten nicht mehr testbar.

In der Praxis werden fast ausschließlich zweiseitige Hypothesen angewendet [6]. Ausnahmen bilden hierbei Nichtüberlegenheitsstudien. Die Entscheidung für oder gegen die Nullhypothese wird mithilfe eines statistischen Signifikanztests[2] herbeigeführt

[2] Dabei wird das Konzept des Signifikanztests nach J. Neyman und E. Pearson zugrunde gelegt. Sie erweitern die Konzepte von R. A. Fisher, in denen lediglich eine Nullhypothese betrachtet wird, um die Alternativhypothese. Zum Fehler 1. Art kommt so noch der Fehler 2. Art hinzu, ohne dessen Berücksichtigung Fallzahlberechnungen und Powerkalkulationen nicht möglich wären.

(siehe z. B. [7]), wobei bei einer Hypothesenformulierung wie in Formel 1 ein zweiseitiger *t*-Test für unabhängige Stichproben zu verwenden ist (siehe z. B. [8]). Die Fehler, die man bei der Entscheidung begehen kann, sind schon in Abb. **1** dargestellt.

Im Folgenden wird anhand von zwei Beispielen die Vorgehensweise bei der Fallzahlkalkulation dargestellt. Dabei betrachten wir nur zweiseitige Hypothesen zum statistischen Nachweis eines Unterschieds. Die dargestellten Grundprinzipien für Fallzahl- und Powerberechnungen gelten aber genauso auch für einseitige Hypothesen. Im ersten Beispiel geht es um die Berechnung des Stichprobenumfangs bei Mittelwertsvergleichen, im zweiten Beispiel um die Berechnung des Stichprobenumfangs beim Vergleich von Proportionen.

Beispiel 1: Fallzahlkalkulation beim Vergleich von zwei Mittelwerten

Dem ersten Beispiel liegt die Frage zugrunde: „Wie viele Probanden müssen in eine Studie eingeschlossen werden, in der die Wirkung einer berufsbegleitenden Nachsorgemaßnahme nach einer kardiologischen Anschlussheilbehandlung in einem parallelen Gruppenvergleich (mit Nachsorge vs. ohne Nachsorge) untersucht werden soll?“

Vor Durchführung der eigentlichen Fallzahlkalkulation ist eine Reihe von Vorüberlegungen nötig. So muss zunächst festgelegt werden, mittels welcher Zielgröße die Wirkung der Nachsorgemaßnahme operationalisiert werden soll. Da hier das Interesse der Berufseingliederung nach der Anschlussheilbehandlung (AHB) gilt, ist eine nahe liegende Zielgröße die Anzahl der Arbeitsunfähigkeitstage (AU-Tage) im Verlauf eines Jahres nach der Anschlussheilbehandlung. Des Weiteren ist zu überlegen, welcher Unterschied in der Wirkung zwischen AHB mit Nachsorge und AHB ohne Nachsorge aus medizinischer und gesellschaftlicher Sicht von Relevanz ist, welcher Unterschied es also „wert“ ist, dass man ihn entdeckt. Diese sog. „relevante Differenz“ hat in diesem Beispiel den Wert von 21 AU-Tagen.[3] Folglich geht man davon aus, dass die entsprechende berufsbegleitende Nachsorge erst dann sinnvoll ist, wenn bei Patienten, die am Nachsorgeprogramm teilnehmen, im Mittel mindestens 21 AU-Tage weniger auftreten als bei Patienten, die keine Nachsorge erhalten.

Im nächsten Schritt wird das Testproblem formuliert, das folgende Gestalt hat:

Formel 5

$$H_0 : \mu_{mN} - \mu_{oN} = 0 \text{ vs. } H_1 : \mu_{mN} - \mu_{oN} \neq 0$$

Dabei bezeichnen μ_{mN} bzw. μ_{oN} die mittlere Anzahl an AU-Tagen der Patienten in der Grundgesamtheit mit bzw. ohne Nachsorge. Festzulegen bleiben noch das Signifikanzniveau und die Power, wobei hier mit $\alpha = 0{,}05$ und $1 - \beta = 0{,}80$ die üblichen Werte gewählt werden.

Als letzte Information vor der Berechnung des Stichprobenumfangs benötigt man noch Angaben zur Variabilität der Zielgröße in Nachsorge- und Vergleichsgruppe. Hier besteht das Problem allerdings darin, dass man über diese Information nicht ohne Weiteres verfügt, ein Aspekt, auf den später noch eingegangen wird. An dieser Stelle sei jedoch vorausgesetzt, dass für beide Gruppen eine Standardabweichung (SD) von 70 AU-Tagen (d. h. $\sigma_{mN} = \sigma_{oN} = 70$) zugrunde gelegt werden kann.

Sämtliche „Zutaten“ zur Berechnung des Stichprobenumfangs sind nun bekannt und können in die entsprechende Formel 6 eingesetzt werden, wobei *n* den erforderlichen Stichprobenumfang je Gruppe (mit Nachsorge oder ohne Nachsorge) beschreibt:

Formel 6

$$n = \frac{\sigma_{mN}^2 + \sigma_{oN}^2}{(\mu_{mN} - \mu_{oN})^2} (z_{1-\alpha/2} + z_{1-\beta})^2 = \frac{70^2 + 70^2}{21^2} (1{,}96 + 0{,}84)^2 = 174$$

Dabei stellen $z_{1-\alpha/2}$ bzw. $z_{1-\beta}$ das $(1 - \alpha/2)$- bzw. $(1 - \beta)$-Quantil der Standardnormalverteilung dar, deren Werte (1,96 für $\alpha = 0{,}05$ und 0,84 für $\beta = 0{,}20$) in entsprechenden Vertafelungen der Standardnormalverteilung (wie z. B. in [9]) abzulesen sind oder mithilfe eines Statistikprogramms berechnet werden können.

Statt Formel 6 lässt sich bei einem Signifikanzniveau von 0,05 und einer Power von 80 % zur etwas gröberen Abschätzung des Stichprobenumfangs auch die vereinfachte Formel 7 verwenden [10]:

Formel 7

$$n = 16 \cdot \frac{\sigma^2}{d^2} = 16 \cdot \frac{4900}{441} = 178$$

Dabei gilt $\sigma = \sigma_{mN} = \sigma_{oN}$ und $d = \mu_{mN} - \mu_{oN}$. Mit beiden Formeln kommt man zu dem Ergebnis, dass insgesamt ca. 350 Probanden benötigt werden, um mit einer Power von 80 % ein signifikantes Testergebnis zu bekommen, wenn in Wahrheit die Differenz im Mittel 21 AU-Tage beträgt.

Eine weit verbreitete Fehlinterpretation ist, dass man mit 350 Probanden nunmehr genügend Power hat, um nachzuweisen, dass zwischen der AHB mit und ohne Nachsorge ein Unterschied in Höhe der klinisch relevanten Differenz von 21 AU-Tagen besteht. Dies ist jedoch nicht der Fall. Das Einsetzen von 21 AU-Tagen ist eine Annahme über einen wahren Effekt, die man benötigt, da die Power eine Funktion des gewählten Effektmaßes ist. Mithilfe eines gewöhnlichen statistischen Signifikanztests lässt sich nur die Nullhypothese ablehnen, dass kein Effekt vorhanden ist. Man hat damit belegt, dass ein Unterschied irgendeiner Größenordnung besteht, aber nicht, dass dieser Unterschied klinisch relevant ist. Zur Begründung klinisch relevanter Unterschiede können später bei der Datenauswertung der Studie Konfidenzintervalle verwendet werden.

[3] Die relevante Differenz von 21 AU-Tagen wurde auf Grundlage der Überlegung gewählt, was eine Rehabilitationsmaßnahme mindestens leisten muss, um als erfolgreich betrachtet werden zu können. Man kam zu dem Schluss, dass dies der Fall ist, wenn die Zeit der Arbeitsunfähigkeit um mindestens drei Wochen reduziert werden kann.

Anzumerken ist noch, dass der *t*-Test, mit dem im Anschluss an diese Fallzahlkalkulation untersucht wird, ob ein signifikanter Unterschied vorliegt, voraussetzt, dass die Daten normalverteilt sind und beide Gruppen gleiche Varianzen aufweisen. Diese Voraussetzungen sollten zumindest annähernd erfüllt sein. Sind die Daten sehr schief verteilt, so hilft oft eine logarithmische Transformation, um eine symmetrische Verteilung der Daten sowie eine Varianzstabilisierung zu erhalten. Für die Fallzahlplanung benötigt man dann natürlich die entsprechenden Angaben (Mittelwertsdifferenz und SD) für die logarithmierten Daten. In unserem Beispiel ist die Annahme, dass AU-Tage normalverteilt sind, aufgrund der oftmals schiefen Verteilung solcher Variablen möglicherweise nicht ohne Weiteres haltbar. Da die errechnete Fallzahl aber deutlich größer als 100 ist, können die auf der Normalverteilung basierenden Formeln 6 und 7 dennoch verwendet werden, da bei großen Stichproben die entsprechenden Berechnungen zumindest asymptotisch richtig sind.

Beispiel 2: Fallzahlkalkulation beim Vergleich von zwei Proportionen

Im zweiten Beispiel geht es um die Berechnung des Stichprobenumfangs beim Vergleich von Proportionen bzw. Anteilen. Die Fragestellung ist hier, wie viele Probandinnen benötigt werden, um zu untersuchen, wie sich „Nordic Walking“ (mit Stockeinsatz) auf den Anteil der Patientinnen auswirkt, die nach erfolgter Brustkrebsoperation Lymphödeme entwickeln. Als Kontrollgruppe dienen Patientinnen, die herkömmliches „Walking“ (ohne Stockeinsatz) betreiben. Medizinisch liegt dieser Fragestellung die Annahme zugrunde, dass sich Nordic Walking anders auf die Bildung von Lymphödemen auswirkt als Walking, da die Arme verstärkt in den Bewegungsablauf integriert werden.

Die Informationen, die zur Berechnung des erforderlichen Stichprobenumfangs benötigt werden, sind die gleichen wie im vorangegangenen Beispiel. Der Unterschied besteht allerdings darin, dass jetzt das Maß für den Unterschied zwischen den Gruppen keine Differenz von Mittelwerten, sondern eine Differenz von Proportionen ist. Außerdem ist hier keine zusätzliche Angabe über die Variabilität der Daten notwendig, weil diese Information implizit in den Proportionen enthalten ist. Die Zielgröße ist die Bildung von Lymphödemen, der zugrunde liegende statistische Test ist ein zweiseitiger Fisher-Test, und für das Signifikanzniveau sowie die Power werden wieder die Werte 0,05 und 0,80 gewählt. Als klinisch relevant wird eine Erhöhung des Anteils um 2,5 % (d. h. $\pi_2 = 0{,}075$) in der Nordic-Walking-Gruppe angesehen, unter der Annahme, dass in der Walking-Gruppe der Anteil 5 % beträgt (d. h. $\pi_1 = 0{,}05$). Diese Angaben werden in Formel 8 eingesetzt, wobei *n* wieder die Anzahl der zu behandelnden Patientinnen je Gruppe bezeichnet:

Formel 8

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})} + z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}\right)^2}{(\pi_1-\pi_2)^2},\quad \bar{\pi} = \frac{\pi_1+\pi_2}{2}$$

$$= \frac{(1{,}96\sqrt{2\cdot 0{,}0625\cdot 0{,}9375} + 0{,}84\sqrt{0{,}05\cdot 0{,}95 + 0{,}075\cdot 0{,}925})^2}{(0{,}075-0{,}05)^2} = \frac{0{,}918}{0{,}000625} = 1469$$

Auch für die Fallzahlkalkulation beim Vergleich von Proportionen gibt es eine Näherungsformel (siehe Formel 9) für ein Signifikanzniveau von 0,05 und eine Power von 0,80 (siehe [10]).

Formel 9

$$n = 16 \cdot \frac{\bar{\pi}(1-\bar{\pi})}{d^2} = 16 \cdot \frac{0{,}0625 \cdot 0{,}9375}{0{,}000625} = 1500$$

Dabei ist $\bar{\pi} = (\pi_1 + \pi_2)/2$ und $d = \pi_1 - \pi_2$. Mit beiden Formeln kommt man also zu dem Ergebnis, dass insgesamt ca. 3000 Probanden benötigt werden, um mit einer Power von 80 % ein signifikantes Testergebnis zu bekommen, wenn in Wahrheit die Differenz der Anteile 2,5 % beträgt. Ein solch hoher Stichprobenumfang ist durchaus typisch bei dichotomen Endpunkten wie „Lymphknotenbildung ja/nein“, da – im Vergleich zu stetigen Zielvariablen – die Power bei dichotomen Endpunkten meist vergleichsweise niedrig ist.

Zusammenhang zwischen Fallzahl, Power, Signifikanzniveau, relevantem Unterschied und Zielgrößenvariabilität

Bei der Fallzahlkalkulation gibt es vier Größen, die veränderlich sind, bzw. vier „Schrauben“, an denen sich „drehen“ lässt. In den hier dargestellten Beispielen waren die „Schrauben“ Power, Signifikanzniveau und klinisch relevanter Unterschied (mit Werten von 0,80, 0,05 und 21 AU-Tage bzw. 2,5 %) bereits „festgedreht“, so dass nur noch zu berechnen war, wie fest die „Schraube“ Fallzahl angezogen werden muss. Allerdings ist es auch denkbar, das Signifikanzniveau, die Fallzahl und den klinisch relevanten Unterschied vorzugeben und auf dieser Grundlage die Power zu berechnen. Ergibt diese Powerberechnung einen Wert von beispielsweise 50 %, so sollte man sich gut überlegen, ob die Durchführung einer Studie unter diesen Bedingungen sinnvoll ist. Die Wahrscheinlichkeit, hier ein signifikantes Ergebnis zu erhalten, wenn in Wahrheit ein Unterschied vorhanden ist, entspricht der Wahrscheinlichkeit, beim Werfen einer Münze „Kopf“ zu erhalten.

Eine weitere Möglichkeit, an den Schrauben zu drehen, besteht darin, Signifikanzniveau, Power und Fallzahl vorzugeben und zu berechnen, wie groß der wahre Effekt sein muss, um ein signifikantes Ergebnis zu erhalten. Allerdings ist hierbei zu beachten, dass u. U. die Fallzahl zu gering ist, um einen Effekt von *klinisch relevanter* Größe entdecken zu können. Des Weiteren sollte überlegt werden, ob aufgrund von Vorwissen (z. B. aus anderen Studien) der Effekt, der notwendig ist, um ein signifikantes Ergebnis zu erhalten, überhaupt realistisch ist. Ist dieser „erwartete Effekt“ sehr viel kleiner als der bei festem Stichprobenumfang berechnete notwendige Effekt bzw. der vorgegebene klinisch relevante Effekt, so ist in dieser Situation trotz Fallzahlkalkulation kein signifikantes Ergebnis zu erwarten.

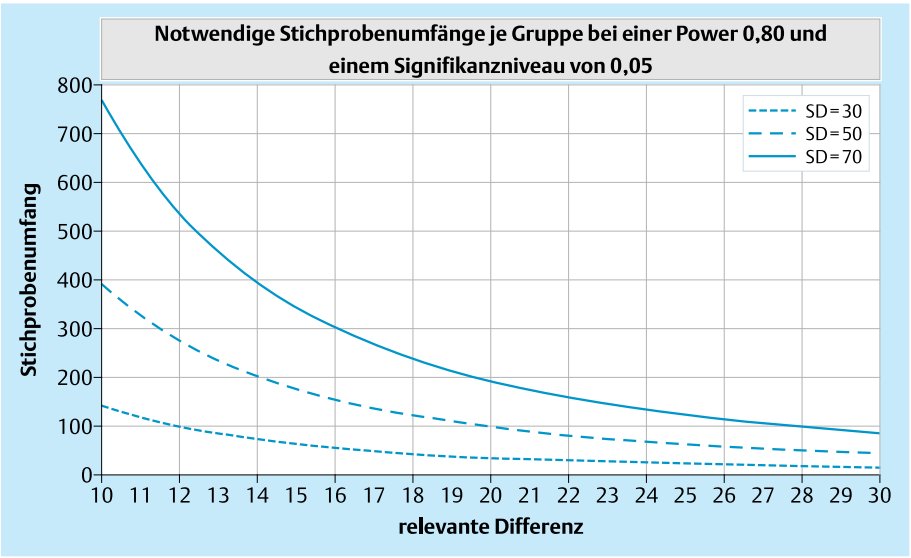

Abb. **2** Notwendige Stichprobenumfänge je Gruppe in Abhängigkeit von Standardabweichung (SD) und relevanter Differenz (für einen zweiseitigen t-Test mit einer Power von 0,80 auf einem Signifikanzniveau von 0,05).

Abb. **2** zeigt, wie sich Veränderungen in der Variabilität der Zielgröße von Beispiel 1 (Standardabweichung in AU-Tagen) und der relevanten Differenz (in AU-Tagen) auf den Stichprobenumfang auswirken. Abb. **2** macht deutlich, dass – bei gleich bleibenden Werten für Signifikanzniveau und Power – der Stichprobenumfang zum einen mit kleiner werdender relevanter Differenz (bei unveränderter Standardabweichung) und zum anderen mit größer werdender Standardabweichung (bei gleich bleibender relevanter Differenz) ansteigt. Dies lässt sich nachvollziehen, wenn man bedenkt, dass es umso schwieriger ist, etwas zu entdecken, je kleiner es ist bzw. je größer der Bereich ist, den es abzusuchen gilt. Ergo muss der Aufwand – hier in Form des Stichprobenumfangs – entsprechend groß sein, sofern die Suche erfolgreich verlaufen soll (vorausgesetzt, dass das, was man sucht, auch tatsächlich vorhanden ist). Des Weiteren sei darauf hingewiesen, dass mit steigendem Signifikanzniveau (bei unveränderter Fallzahl und gleich bleibendem klinisch relevanten Unterschied) die Power größer wird und umgekehrt.

Konsequenzen einer zu kleinen Fallzahl

Beaupre et al. [11] untersuchten den Einfluss eines präoperativen Trainingsprogramms auf die Wiedererlangung der Kniefunktionalität, die gesundheitsbezogene Lebensqualität und die Inanspruchnahme von Gesundheitsdienstleistungen nach vollständiger Kniearthroplastie. Die Berechnung des Stichprobenumfangs erfolgte bezüglich eines Indexes zur Einschätzung der Funktionalität des Knies auf einem Signifikanzniveau von 0,05 und einer Power von 0,80. Es ergab sich eine Fallzahl von insgesamt 131 Patienten, von denen 66 in die Interventionsgruppe (mit Trainingsprogramm) und 65 in die Kontrollgruppe (ohne Trainingsprogramm) randomisiert wurden.

Hinsichtlich einer weiteren relevanten Zielvariablen, der „Aufenthaltsdauer im Krankenhaus“, ließ sich kein signifikanter Unterschied zwischen Interventions- und Kontrollgruppe nachweisen. Beaupre et al. führen dies u. a. darauf zurück, dass ihre Studie in Bezug auf diese Zielvariable eine zu geringe Power aufweist. Wie groß (bzw. klein) die Power tatsächlich ist, lässt sich ermitteln, wenn man die Formel zur Fallzahlberechnung zunächst nach $z_{1-\beta}$ auflöst und dann die aus der Studie von Beaupre et al. bekannten Werte entsprechend einsetzt:

Formel 10

$$n = \frac{\sigma_{mT}^2 + \sigma_{oT}^2}{(\mu_{mT} - \mu_{oT})^2} (z_{1-\alpha/2} + z_{1-\beta})^2 \Leftrightarrow z_{1-\beta} = \sqrt{n \frac{(\mu_{mT} - \mu_{oT})^2}{\sigma_{mT}^2 + \sigma_{oT}^2}} - z_{1-\alpha/2}$$

Da laut Beaupre et al. ein Unterschied zwischen Interventions- und Kontrollgruppe von 2 Tagen Aufenthaltsdauer (im Mittel) als relevant angesehen wird, gilt $\mu_{mT} - \mu_{oT} = 2$ (mT = „mit Training“, oT = „ohne Training“). Für die Varianzen werden die Werte verwendet, die sich in der Studie ermitteln ließen: $\sigma_{mT}^2 = 20{,}25$ und $\sigma_{oT}^2 = 27{,}04$. Da das Signifikanzniveau 0,05 beträgt, hat $z_{1-\alpha/2}$ wieder den Wert 1,96. Mit $n = 65$ ergibt sich damit:

Formel 11

$$z_{1-\beta} = \sqrt{n \frac{(\mu_{mT} - \mu_{oT})^2}{\sigma_{mT}^2 + \sigma_{oT}^2}} - z_{1-\alpha/2} = \sqrt{65 \frac{(2)^2}{20{,}25 + 27{,}04}} - 1{,}96 = 0{,}38$$

In Vertafelungen der Standardnormalverteilung (siehe z. B. [9]) ist dann nachzuschlagen, dass für $z_{1-\beta} = 0{,}38$ die Power einen Wert von ca. 0,65 aufweist. Die Wahrscheinlichkeit, ein signifikantes Resultat zu erhalten, wenn in Wahrheit ein mittlerer Unterschied von 2 Tagen vorhanden ist, beträgt hier bei einem Stichprobenumfang von $n = 131$ also nur 65 %.

Beaupre et al. halten den Parameter „Aufenthaltsdauer“ allerdings für so wichtig, dass sie empfehlen, ihn in zukünftigen Studien als Hauptzielgröße zu betrachten. Bei der entsprechenden Fallzahlkalkulation würde man dann – wie üblich – eine Power von 80 % verwenden. Der relevante mittlere Unterschied beträgt 2 Tage Aufenthaltsdauer, und für die Varianzen werden wieder die obigen Werte verwendet. Damit erhält man:

Formel 12

$$n = \frac{\sigma_{mT}^2 + \sigma_{oT}^2}{(\mu_{mT} - \mu_{oT})^2} (z_{1-\alpha/2} + z_{1-\beta})^2 = \frac{20{,}25 + 27{,}04}{(2)^2} (1{,}96 + 0{,}84)^2 = 93$$

Es wären also 93 Patienten je Gruppe notwendig, um mit einer Wahrscheinlichkeit von 80% ein signifikantes Resultat zu erhalten, wenn in Wahrheit ein mittlerer Unterschied von 2 Tagen vorhanden ist.

Variabilität der Zielgröße

Wie schon erwähnt, ist beim Vergleich von Mittelwerten eine zur Berechnung des benötigten Stichprobenumfangs wesentliche Information zu Beginn der Studie nicht bekannt: die Variabilität der Zielgröße in Interventions- und Vergleichsgruppe. Ein möglicher Ausweg aus diesem Dilemma ist die Schätzung der Variabilität anhand von bereits durchgeführten Studien, die die gleiche Zielgröße haben. Ein Problem ist hierbei allerdings darin zu sehen, dass Unterschiede in den Designmerkmalen zwischen früherer und aktueller Studie (andere Ein-/Ausschlusskriterien, unterschiedliche Zahl von Zentren etc.) bestehen können.

Ein anderer Ausweg besteht darin, zur Schätzung der Variabilität eine Pilotstudie durchzuführen. Hier gibt es zwei Möglichkeiten. Die eine ist die, die Pilotstudie unter den Bedingungen des Studienprotokolls der eigentlichen Hauptstudie vorzuschalten. Ein Nachteil ist allerdings darin zu sehen, dass die in die Pilotstudie eingeschlossenen Patienten ausschließlich der Schätzung der Variabilität der Zielgröße dienen und nicht auch schon der Schätzung des Behandlungseffekts. Hier werden also Informationen verschenkt. Insofern besteht die zweite – etwas elegantere – Möglichkeit darin, die Pilotstudie als Bestandteil der Hauptstudie zu betrachten. Hierbei wird zunächst zur initialen Fallzahlberechnung die Variabilität aus früheren Studien verwendet. Nachdem ein Teil der Patienten – üblich sind Werte zwischen 25% und 50% – die Studie durchlaufen hat, wird die Variabilität anhand dieser Patienten geschätzt und die Fallzahl erneut berechnet. Allerdings dürfen bei gewöhnlichen Designs Änderungen im Design oder im Studienprotokoll nicht mithilfe der Daten aus der Pilotstudie vorgenommen werden, wenn man diese als Bestandteil der Hauptstudie verwenden will. Mehr Flexibilität besitzen gruppensequenzielle Pläne und insbesondere adaptive Designs [12], auf die wir im Rahmen dieses Artikels jedoch nicht weiter eingehen können.

Eine weitere, in der Praxis häufig angewendete Vorgehensweise besteht darin, die relevante Differenz d nicht absolut, sondern relativ – bezogen auf die Standardabweichung σ – zu definieren, d. h., man betrachtet die *standardisierte* Differenz d/σ. Auf diese Weise lässt sich z. B. eine relevante Differenz in der Größenordnung einer halben Standardabweichung definieren, und es ist nicht erforderlich, explizite Werte für die Differenz und die Streuung anzugeben. In der Literatur werden Richtwerte der standardisierten Differenz für „kleine", „mittlere" und „große" Gruppenunterschiede angegeben [13]. Dieses scheinbar verlockende Vorgehen sollte allerdings nur dann gewählt werden, wenn die Festlegung eines relevanten Unterschieds a priori nicht möglich ist (was z. B. bei bislang unbekannten Effekten der Fall sein kann). Denn ein gravierender Nachteil ist darin zu sehen, dass eine für die Interpretation wichtige Festlegung eines *inhaltlich relevanten* Unterschiedes auf diese Weise nicht vorgenommen wird.

Diskussion

In den Rehabilitationswissenschaften – wie auch in anderen Bereichen der medizinischen und sozialwissenschaftlichen Forschung – sind vor der Durchführung von Studien solide Fallzahlkalkulationen unerlässlich. Nur auf diese Weise lässt sich vermeiden, dass zu wenige oder auch zu viele Probanden in die Studie eingeschlossen werden. Ersteres ist u. a. dann fatal, wenn aufgrund eines zu geringen Stichprobenumfangs ein eigentlich vorhandener relevanter Effekt übersehen wird. Letzteres ist u. a. aus ethischen Gründen bedenklich, da in diesem Fall unnötig viele Patienten einer nicht wirksamen oder möglicherweise auch schädlichen Therapie ausgesetzt werden.

Das Konzept der Fallzahl- und Powerberechnung ist primär sinnvoll *vor* Durchführung einer Studie. Das in der Praxis häufig beobachtete Vorgehen, *nach* Durchführung der Studie Aussagen basierend auf der Power des *beobachteten Unterschiedes* zu treffen, ist irreführend. Diese sog. „beobachtete Power" ist eine Funktion des p-Wertes des durchgeführten Tests und enthält somit keine zusätzlichen Informationen. Aufgrund des mathematischen Zusammenhangs ist vorgegeben, dass die beobachtete Power mit steigendem p-Wert sinkt. Eine nachträgliche Begründung eines nichtsignifikanten Unterschieds „aufgrund" zu geringer Power ist in diesem Fall sinnlos, da ein hoher p-Wert *immer* mit einer geringen beobachteten Power verbunden ist [14].

Schon eher sinnvoll kann dagegen eine nach Durchführung der Studie vorgenommene Powerberechnung sein, wenn diese nicht auf dem beobachteten Unterschied, sondern auf dem als *klinisch relevant* erachteten Unterschied basiert. Eine solche Analyse kann zeigen, dass bei dem verwendeten Stichprobenumfang die Power für ein signifikantes Testergebnis zu gering war (siehe Abschnitt „Konsequenzen einer zu geringen Fallzahl"). Noch umfassendere Information liefert in solchen Situationen jedoch das Konfidenzintervall für das gewählte Effektmaß. Ein Konfidenzintervall beschreibt, welche Effektstärken mit den beobachteten Daten verträglich sind [15]. Diese Information ist klinisch interpretierbar und hat somit Vorteile gegenüber der rein statistischen Information einer Powerangabe.

In jüngster Zeit wurde kritisiert, dass die „Beschwörung der adäquaten Power die Diskussion über andere methodische Fragen" überdecke. Dies sei umso bedenklicher, da die Fallzahlkalkulation – z. B. aufgrund unzureichender Kenntnis der Zielgrößenvariabilität – niemals vollständig exakt sein könne. Beispielsweise führe ein unzureichendes Randomisierungsverfahren zu verzerrten Studienergebnissen, die auch dann nicht mehr zu „retten" sind, wenn eine große Stichprobe eine hohe Genauigkeit verspricht [16]. Diese Kritik ist zweifelsohne berechtigt. Es darf nicht übersehen werden, dass die Fallzahlkalkulation nur *ein* Element einer sorgfältigen Studienplanung darstellt: Sie ist zwar notwendige, aber nicht hinreichende Bedingung für eine hohe methodische Qualität jeder Studie. Aus unserer Sicht ist es dennoch erforderlich, Forscher nicht nur z. B. hinsichtlich einer korrekten Randomisierung, sondern auch in Bezug auf eine adäquate Fallzahlplanung in die Pflicht zu nehmen. Denn bei aller Unsicherheit, mit der sie behaftet ist, führt sie doch zumindest dazu, dass der Wissenschaftler „gezwungen" wird, sich über zentrale Aspekte des Forschungsprozesses gründlich Gedanken zu ma-

chen. Dazu gehören eine exakte Spezifizierung und Formulierung der Hauptfragestellung, die geeignete Operationalisierung der entsprechenden Zielgröße, Wahl des geeigneten Testverfahrens sowie Überlegungen zur klinischen Relevanz möglicher Effekte.

Fallzahlkalkulationen können (und müssen) nicht nur für die hier beispielhaft vorgestellten – noch recht unkomplizierten – Designs durchgeführt werden. So können erforderliche Stichprobenumfänge z. B. auch bei Überlebenszeitanalysen, Nichtunterlegenheits- und Äquivalenzfragestellungen sowie nichtparametrischen Verfahren berechnet werden. Hierzu sei jedoch auf die weiterführende Literatur (z. B. [1,17,18] und [19]) verwiesen.

Zur praktischen Durchführung von Stichproben- und Powerberechnungen gibt es eine Reihe von kommerziellen und frei verfügbaren Softwarepaketen. Eine vergleichende Übersicht mit genauer Beschreibung der jeweiligen Möglichkeiten geben Ortseifen et al. [20]. Eine sehr häufig verwendete kommerzielle Software ist z. B. das Windows-Programm nQuery, das mittlerweile als Version 6.0 erhältlich ist (http://www.statsol.ie/nquery/nquery.htm). Neben den frei verfügbaren Programmen gibt es auch entsprechende webbasierte Tools, z. B. den „Power Calculator", der von der Abteilung Statistik der Universität von Kalifornien zur Verfügung gestellt wird (http://calculators.stat.ucla.edu/powercalc/).

Dank

Wir bedanken uns bei den internen Gutachtern Prof. Dr. Dr. H. Faller (Würzburg), Prof. Dr. Th. Kohlmann (Greifswald) und Dr. Ch. Zwingmann (Siegburg) sowie zwei uns unbekannten externen Gutachtern für zahlreiche wertvolle Hinweise.

Literatur

[1] Bock J. Bestimmung des Stichprobenumfangs für biologische Experimente und kontrollierte klinische Studien. München: Oldenbourg, 1998: 1
[2] Beck-Bornholdt HP, Dubben HH. Der Hund der Eier legt – Erkennen von Fehlinformationen durch Querdenken. Reinbek: Rowohlt, 2001: 113 – 114
[3] Dubben HH, Beck-Bornholdt HP. Was ist Power und warum ausgerechnet 80 %? Medizinische Klinik 1999; 94, Suppl II: 5 – 7
[4] Faller H. Signifikanz, Effektstärke und Konfidenzintervall. Rehabilitation 2004; 43: 174 – 178
[5] Knottnerus JA, Bouter LM. The ethics of sample size: Two-sided testing and one-sided thinking. Journal of Clinical Epidemiololgy 2001; 54: 109 – 110
[6] Bland JM, Altman DG. One and two sided tests of significance. British Medical Journal 1994; 309: 248
[7] Lange S, Bender R. Was ist ein Signifikanztest? – Allgemeine Aspekte. Deutsche Medizinische Wochenschrift 2001; 126: T42 – T44
[8] Bender R, Lange S, Ziegler A. Wichtige Signifikanztests. Deutsche Medizinische Wochenschrift 2002; 127: T1 – T3
[9] Hartung J. Statistik – Lehr- und Handbuch der angewandten Statistik. München: Oldenbourg, 1991: 891
[10] Lehr R. Sixteen s-squared over d-squared: A relation for crude sample size estimates. Statistics in Medicine 1992; 11: 1099 – 1102
[11] Beaupre LA, Lier D, Davies DM, Johnston DBC. The effect of a preoperative exercise and education program on functional recovery, health related quality of life, and health service utilization following primary total knee arthroplasty. Journal of Rheumatology 2004; 31 (6): 1166 – 1173
[12] Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics 1999; 55: 1286 – 1290
[13] Cohen J. A power primer. Psychological Bulletin 1992; 112: 155 – 159
[14] Lenth RV. Some practical guidelines for effective sample size determination. American Statistician 2001; 55: 187 – 193
[15] Hoening JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. American Statistician 2001; 55: 19 – 24
[16] Schulz KF, Grimes DA. Fallzahlschätzung in randomisierten Studien: ein Muss und ein Mysterium. Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen 2006; 100: 129 – 135
[17] Bock J, Toutenburg H. Sample size determination in clinical research. In: Rao CR, Chakraborty R (eds): Handbook of Statistics (Vol. 8). Amsterdam: Elsevier, 1991: 515 – 538
[18] Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials 1981; 2: 93 – 113
[19] Julious SA. Sample sizes for clinical trials with normal data. Statistics in Medicine 2004; 23: 1921 – 1986
[20] Ortseifen C, Bruckner T, Burke M, Kieser M. An overview of software tools for sample size determination. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 1997; 28: 91 – 118