# Epidemiologie 1

## Fallzahlschätzung in randomisierten Studien: ein Muss und ein Mysterium

Kenneth F. Schulz und David A. Grimes
Family Health International, PO Box 13950, Research, Triangle Park, NC 27709, USA

aus: *Lancet* 2005; **365**: 1348–53: Sample size calculation in randomized trials: mandatory and mystical


### Zusammenfassung

Vor Beginn einer randomisierten Studie sollte der Untersucher die erforderliche Stichprobengröße korrekt berechnen und die Einzelheiten dazu im zugehörigen Studienbericht niederlegen. In diese vorab durchgeführten Berechnungen fließen bei der Festlegung der zu entdeckenden Effektgröße – z. B. Ereignisraten in Behandlungs- und Kontrollgruppen – bereits subjektive klinische Entscheidungen ein. Überdies haben diese Entscheidungen einen starken Einfluss auf die Berechnung der Stichprobengrößen. Wir halten es für zweifelhaft, Studien auf der Grundlage ungenau berechneter Stichprobengrößen als unethisch zu brandmarken. Vielmehr könnte man Studien mit zu kleinen Stichproben und damit ungenügender statistischer Trennschärfe oder Power – in diesem Fall spricht man von „underpowered" – akzeptieren, wenn sich die Untersucher mit methodischer Strenge bemühen, systematische Fehler oder Verzerrungen (sog. Bias) auszuschalten, wenn sie korrekte Studienberichte abfassen, um Missverständnissen vorzubeugen, und alle ihre Ergebnisse publizieren, um dem Vorwurf des Publikationsbias zu begegnen. Eine Verlagerung des Schwerpunktes von der Stichprobengröße hin zur methodischen Qualität würde uns mehr Studien mit weniger Biasproblemen bescheren. Unverzerrte Studien mit unpräzisen Ergebnissen sind immer noch besser als gar keine. Ärzten und Patienten sind wir in diesem Punkt Aufklärung schuldig.


Die Forderung nach der Berechnung von Stichprobengrößen für randomisierte Studien scheint unangreifbar zu sein. In der Tat sollten Untersucher die Größe ihrer Stichproben korrekt berechnen und die wichtigsten Details dazu in ihrem Studienbericht niederlegen. Wissenschaftsmethodiker beschreiben entsprechende Vorgehensweisen in Buch- und Zeitschriftenbeiträgen, Protokoll- und Ethikkommissionen verlangen deren Einhaltung, und die CONSORT-Leitlinien für die Abfassung von Studienberichten enthalten genaue Vorgaben für die Darlegung von Fallzahlschätzungen [1, 2]. In diesem Punkt sind sich fast alle einig.

Mit der Veröffentlichung der Arbeiten einer Forschergruppe unter der Leitung von Tom Chalmers geriet diese Einmütigkeit der medizinischen Welt vor mehr als einem Vierteljahrhundert jäh ins Wanken. In ihrem zum Meilenstein gewordenen Artikel gingen die Autoren nämlich auf den Mangel an statistischer Power bei so genannten negativen randomisierten Studien ein, die in führenden allgemeinmedizinischen Fachzeitschriften veröffentlicht worden waren [3]. Während seiner langen glänzenden Karriere hat Chalmers Hunderte von Artikeln publiziert. Seine Abhandlung über Stichprobengröße und statistische Power ist vielfach zitiert worden, was ihn paradoxerweise sehr beunruhigt hat [4]. Denn seiner Meinung nach hat diese Veröffentlichung mehr Schaden angerichtet als alle anderen Artikel, an denen er jemals beteiligt war. Wieso? Seine Gründe werden wir später erläutern – also gedulden Sie sich noch einen Moment.

## Elemente der Fallzahlschätzung

Zur Berechnung der Stichprobengröße für Studien mit dichotomen Endpunkten (z. B. krank *vs.* gesund) benötigt man vier Größen: den Typ I-Fehler (a), die statistische Trennschärfe (Power), die Ereignisrate in der Kontrollgruppe und den interessierenden Therapieeffekt (oder eine entsprechende Ereignisrate in der Behandlungsgruppe). Diese elementaren Bestandteile tauchen – außer wenn vielleicht andere Voraussetzungen erforderlich sind – auch bei der Berechnung anderer Arten von Zielparametern (Outcomes) immer wieder auf. Bei quantitativen Outcomes und einem typischen statistischen Test benötigt man Annahmen über den Unterschied zwischen den Mittelwerten sowie die Varianz dieses Unterschieds.

In der klinischen Forschung besteht bei der Prüfung von Hypothesen das Risiko, einen von zwei Fehlern grundsätzlicher Art zu begehen (Kasten 1). Erstens kann es vorkommen, dass der Forscher zu dem Schluss kommt, zwei Therapien seien voneinander verschieden, wenn dies in Wirklichkeit gar nicht der Fall ist. Dieser sog. Typ I- oder $\alpha$-Fehler (Fehler der 1. Art) misst die Wahrscheinlichkeit,

mit der eine solche falsch-positive Schlussfolgerung gezogen wird. Üblicherweise hat sich für $\alpha$ ein Wert von 0,05 eingebürgert. Das bedeutet, der Untersucher möchte, dass die Wahrscheinlichkeit einer falsch-positiven Schlussfolgerung für ihn unter 5% liegt. Zweitens könnte der Forscher folgern, dass zwei Behandlungen sich nicht unterscheiden, wenn sie es tatsächlich aber doch tun – ein sog. falsch-negativer Schluss. Dieser Typ II- oder $\beta$-Fehler misst die Wahrscheinlichkeit einer solchen falsch-negativen Schlussfolgerung. Konventionsgemäß wird für $\beta$ ein Wert von 0,20 angenommen. Das heißt, die Irrtumswahrscheinlichkeit für eine falsch-negative Schlussfolgerung soll weniger als 20% betragen.

Die statistische Power leitet sich vom $\beta$-Fehler ab. Mathematisch betrachtet gilt sie als Komplement zu $\beta$ (also $1 - \beta$) und repräsentiert die Wahrscheinlichkeit, mit der eine falsch-negative Schlussfolgerung vermieden wird. Angenommen, $\beta$ betrüge 0,20. Dann läge die Power bei 0,80 oder 80%. Anders gesagt, die statistische Power gibt die Wahrscheinlichkeit an, mit der ein Unterschied nachgewiesen werden kann (vorausgesetzt, es besteht tatsächlich ein Unterschied der angenommenen Größe, dann gilt ein solcher Unterschied für $p < \alpha$ als signifikant). So liegt z. B. in einer Studie mit einer Power von 0,80 die Wahrscheinlichkeit, einen Unterschied zwischen zwei Therapien aufzudecken, bei 80%, sofern im Untersuchungskollektiv tatsächlich ein Unterschied der vermuteten Größenordnung existiert.

$\alpha$-Fehler, $\beta$-Fehler und Power verstehen zu wollen kann sich zugegebenermaßen als wahre Herausforderung erweisen. Allerdings gibt es in der Regel Konventionen, die dem Untersucher sagen, welche Größen in die Berechnung der Stichprobe einfließen müssen. Die anderen Inputdaten verursachen zwar weniger konzeptuelle, dafür aber mehr praktische Probleme. Als Startwerte muss der Untersucher die wahren Ereignisraten in seinen Behandlungs- und Kontrollgruppen schätzen. Normalerweise empfehlen wir, erst die Ereignisrate im Untersuchungskollektiv zu schätzen und dann den interessierenden Therapieeffekt festzulegen. Angenommen, der Untersucher erwartet für seine Kontrollgruppe beispielsweise eine Ereignisrate von 10%. Dann würde er für die Behandlungsgruppe eine absolute Änderung (z. B. eine absolute Reduktion um 3%), eine relative Änderung (eine relative Reduktion um 30%) oder einfach eine Ereignisrate von 7% schätzen. Von diesen Annahmen ausgehend berechnet der Untersucher seine Stichprobengröße. Diese Verfahren werden in Standardtexten beschrieben, z. B. für binäre und kontinuierliche Zielgrößen sowie Ereigniszeiten [5–7]. Meist werden zur Berechnung von Stichprobengröße und statistischer Power entsprechende Software-Programme benutzt (vorzugsweise unter Anleitung eines Statistikers). Denn bei der Berechnung von Hand stößt man, selbst bei den einfachsten Gleichungen wie etwa in Kasten 2, leider teuflisch schnell an seine Grenzen.

**Kasten 1.** Fehlerdefinition

**Typ I- oder $\alpha$-Fehler**
Die Wahrscheinlichkeit, einen statistisch signifikanten Unterschied zu entdecken, wenn die verglichenen Therapien in Wirklichkeit gleich wirksam sind, z. B. die Chance, ein falsch-positives Ergebnis zu erzielen.

**Typ II- oder $\beta$-Fehler**
Die Wahrscheinlichkeit, einen statistisch signifikanten Unterschied nicht zu entdecken, wenn in Wahrheit ein Unterschied einer bestimmten Größe existiert, z. e. die Chance, ein falsch-negatives Ergebnis zu erzielen.

**Power ($1 - \beta$)**
Die Wahrscheinlichkeit, einen statistisch signifikanten Unterschied nachzuweisen, wenn tatsächlich ein Unterschied einer bestimmten Größe vorliegt.

**Kasten 2.** Die einfachsten Gleichungen zur Berechnung der ungefähren Stichprobengröße für binäre Endpunkte, ausgehend von $\alpha = 0{,}05$, Power = 0,90 und gleich großem Stichprobenumfang in beiden Gruppen

$n$ = Stichprobengröße in jeder der beiden Gruppen
$p_1$ = Ereignisrate in der Behandlungsgruppe (nicht in der Gleichung, sondern implizit in der Schätzung von R und $p_2$ enthalten)
$p_2$ = Ereignisrate in der Kontrollgruppe
$R$ = Relatives Risiko ($p_1/p_2$)

$$n = \frac{10{,}51\,[(R+1) - p_2\,(R^2+1)]}{p_2\,(1-R)^2}$$

Angenommen, wir gehen für die Kontrollgruppe von einer Ereignisrate von 10% aus ($p_2 = 0{,}10$) und legen fest, dass es sich – ausgehend von $\alpha = 0{,}05$ und Power = 0,90 – bei dem klinisch relevanten Unterschied, den es aufzudecken gilt, um eine durch die neue Therapie bewirkte 40%ige Reduktion (R = 0,60) handelt. (Beachte: R = 0,60 entspricht einer Ereignisrate in der Behandlungsgruppe von $p_1 = 0{,}06$, also R = 6%/10%).

$n = 961{,}665 \cong 962$ pro Gruppe (Die PASS-Software, Version 6.0; NCSS, Kaysville, UT, USA, die mit einer genaueren Gleichung arbeitet, errechnet 965.)

Für $\alpha$ und Power können sich auch andere Werte ergeben, wenn man statt 10,51 einen entsprechenden Wert aus der folgenden Tabelle in die Gleichung einsetzt:

| **Power ($1 - \beta$)** | | 0,80 | 0,90 | 0,95 |
|---|---|---|---|---|
| **$\alpha$-Fehler** | 0,05 | 7,85 | 10,51 | 13,00 |
| | 0,01 | 11,68 | 14,88 | 17,82 |

## Wie wirken sich $\alpha$-Fehler und Power aus?

Normalerweise reichen die per Konvention festgelegten Werte von $\alpha = 0{,}05$ (5%) und Power = 0,80 (80%) völlig aus. Je nach Untersuchungsgebiet können aber auch andere Annahmen sinn-


Z. ärztl. Fortbild. Qual. Gesundh.wes. (2006) 100; 129–135
http://www.elsevier.de/zaefq

voll sein. Wenn etwa ein bei Hysterektomien zur Prophylaxe verabreichtes Standardantibiotikum wirksam ist und nur geringe Nebenwirkungen aufweist, könnten wir in einer Studie über ein neues Antibiotikum für $\alpha$ einen niedrigeren Wert (z. B. 0,01) annehmen, um die Wahrscheinlichkeit einer falsch-positiven Schlussfolgerung zu verringern. Wir könnten sogar in Erwägung ziehen, die Power auf unter 0,80 zu senken, da wir weniger Bedenken haben müssen, eine wirksame Therapie zu übersehen. Denn eine wirksame und sichere Behandlung gibt es ja bereits. Dagegen ändert sich die Gewichtung, wenn der Untersucher im gleichen Setting ein Standardantibiotikum im Vergleich zu einer preiswerten sicheren Vitaminergänzung testen will. In dieser Situation würde ein $\alpha$-Fehler kaum Schaden verursachen, sodass die Festlegung auf 0,10 angebracht sein könnte [7]. Würde diese billige und einfache Intervention allerdings von Nutzen sein, würden wir ihren Effekt nicht übersehen wollen. Aus diesem Grund könnte der Untersucher die Power auf 0,99 erhöhen wollen.

Die unterschiedlichen Annahmen für $\alpha$-Fehler und Power wirken sich direkt auf die Stichprobengröße aus. Eine Verringerung von $\alpha$ und eine Erhöhung der Power führen beide zu einer Vergrößerung der Stichprobe: Wenn wir z. B. $\alpha$ von 0,05 auf 0,01 senken, wächst die erforderliche Studiengröße bei einer Power von 0,50 um 70% an, bei einer Power von 0,80 um 50% (s. Tabelle). Bei $\alpha = 0{,}05$ führt die Erhöhung der Power von 0,50 auf 0,80 zu einem zweifachen Anstieg der Studiengröße und bei einer Erhöhung von 0,50 auf 0,99 zu einem nahezu 5fachen Anstieg (s. Tabelle). Je nachdem, welche Werte wir für $\alpha$ und Power zugrunde legen, ergeben sich unterschiedliche Stichprobenumfänge und damit auch unterschiedlich hohe Studienkosten.

Um die geschätzten Stichprobenumfänge zu reduzieren, benutzen manche Untersucher für den $\alpha$-Fehler einseitige Tests. Wir raten von einem solchen Vorgehen ab. Wir haben zwar bislang zweiseitige Tests angewendet, doch bei entsprechenden biologischen Kenntnissen können auch einseitige Tests sinnvoll sein. Diese Entscheidung sollte aber keinen Einfluss auf die Bestimmung der Stichprobengröße haben. Wir empfehlen, denselben Evidenzstandard anzulegen, ganz gleich, ob man von einer ein- oder zweiseitigen Hypothese ausgeht [7]. Ein einseitiger Test mit $\alpha = 0{,}025$ ergibt denselben Evidenzgrad wie ein zweiseitiger Test mit $\alpha = 0{,}05$. Für Stichprobenberechnungen einen einseitigen Test zu benutzen, um die erforderlichen Stichprobenumfänge zu reduzieren, hieße die Gutgläubigkeit des Lesers überstrapazieren.

## Schätzung von Populationsparametern

Für manche Untersucher ist die Schätzung von Populationsparametern – z. B. den Ereignisraten in Behandlungs- und Kontrollgruppen – mit einer Aura des Mysteriösen umgeben. Andere fangen bei dieser Vorstellung an zu spotten, denn die Schätzung dieser Parameter ist schließlich das Ziel der Untersuchung: Dies bereits vor Beginn der Studie zu tun, erscheint ihnen absurd. Der entscheidende Punkt ist hier jedoch, dass diese Forscher nicht die Populationsparameter an sich schätzen, sondern den Behandlungseffekt, den es sich ihrer Meinung nach nachzuweisen lohnt. Das ist ein großer Unterschied!

In der Regel beginnt der Untersucher mit der Schätzung der Ereignisrate in der Kontrollgruppe. Gelegentlich ergeben sich wegen allzu spärlicher Daten nur unzuverlässige Schätzer. So mussten wir beispielsweise für eine Familienplanungsstudie in Nairobi (Kenia) bei Anwenderinnen von Intrauterinpessaren die Ereignisrate für entzündliche Erkrankungen der Organe des kleinen Beckens (PID) schätzen. Behördenmitarbeiter gingen von 40% aus; die Ärzte des medizinischen Versorgungszentrums hielten diesen Wert aber für viel zu hoch und schlugen 12% vor, wir gingen von einer konservativen Schätzung von 6% aus. Der Wert in der Placebogruppe der tatsächlichen randomisierten Studie lag jedoch bei 1,9% [8]. Der erste Schätzer wich also um das mehr als 20fache vom tatsächlichen Wert ab! Und so etwas wirkt sich ganz enorm auf die Stichprobenberechnung aus.

Anhaltspunkte für den Schätzer des Endpunkts in der Kontrollgruppe ergeben sich gelegentlich aus veröffentlichten Berichten. Doch auch wenn sie in der Regel jede Menge Unterschiede enthalten, etwa im Hinblick auf Studienorte, Eignungskriterien, Endpunkte und Therapien, liegen meist bestimmte Informationen über die Kontrollgruppe vor, die uns als Ausgangspunkt dienen können.

In einer Präventionsstudie über Fieber nach Hysterektomie zeigen die als einigermaßen zufrieden stellend geltenden Daten, dass die febrile Morbidität nach prophylaktischer Gabe des Standardantibiotikums bei 10% liegt. Diesen Wert nehmen wir als Ereignisrate in der Kontrollgruppe. Der Schätzer der interessierenden Effektgröße sollte sowohl klinischen Scharfsinn als auch den potenziellen Public-Health-Effekt widerspiegeln – ein wichtiger Aspekt, den man nicht dem Statistiker allein überlassen sollte. Der Entscheidungsprozess setzt sich mit der Sammlung von Informationen über den klinischen Hintergrund fort. Angenommen, die Prophylaxe mit dem Standardantibiotikum kostet US-$ 10, verursacht nur wenige Nebenwirkungen und kann oral verabreicht werden. Die Prophylaxe mit dem neuen Antibiotikum kostet US-$ 200, hat mehr Nebenwirkungen und wird intravenös gegeben, verfügt aber über ein breiteres Wirkspektrum. Alle diese pragmatischen und klinischen Faktoren wirken sich auf den Entscheidungsprozess aus. Würden wir nun in Anbetracht

**Tabelle.** Ungefähre relative Studiengröße bei unterschiedlichen Annahmen für $\alpha$ und Power.

| **Power (1 – $\beta$)** | | 0,50 | 0,80 | 0,90 | 0,99 |
|---|---|---|---|---|---|
| **$\alpha$-Fehler** | 0,05 | 100 | 200 | 270 | 480 |
| | 0,01 | 170 | 300 | 390 | 630 |
| | 0,001 | 280 | 440 | 540 | 820 |

der Ereignisrate (Fieber) von 10% in der Kontrollgruppe und unter Berücksichtigung des klinischen Hintergrundes eine 10%ige Reduktion auf 9% nachweisen wollen, eine 20%ige Reduktion auf 8%, eine 30%ige Reduktion auf 7%, eine 40%ige Reduktion auf 6%, eine 50%ige Reduktion auf 5% usw.? Die Festlegung dieses nachzuweisenden Unterschiedes enthält nämlich bereits eine subjektive klinische Entscheidung, und eine richtige Antwort auf diese Frage gibt es nicht: Auch wenn wir vielleicht der Meinung wären, dass sich eine 30%ige Reduktion nachzuweisen lohnt, würden andere Untersucher sich vielleicht eher für eine Senkung um 50% entscheiden wollen.

Von welchen Parametern wir letztlich ausgehen, hat einen enormen Einfluss auf die Stichprobenberechnung. Wenn wir die Annahmen für die Kontrollgruppe konstant halten, macht die Halbierung der Effektgröße eine mehr als 4fache Zunahme der Studiengröße erforderlich. Entsprechend steigt die Studiengröße um das mehr als 16fache, wenn wir die Effektgröße durch 4 teilen. Anders ausgedrückt, die Stichprobengröße steigt im umgekehrten Verhältnis zum Quadrat der Effektgrößenverringerung (was Statistiker eine quadratische Beziehung nennen). Ausgehend von unseren Eingangsparametern von 10% in der Kontrollgruppe und 6% in der Interventionsgruppe sowie $\alpha = 0{,}05$ und einer Power von 0,90 würden wir pro Gruppe etwa 965 Studienteilnehmer benötigen (Kasten 2). Die Halbierung der Effektgröße und damit die Veränderung des Schätzers für die Interventionsgruppe auf 8% macht eine mehr als 4fache Steigerung des Stichprobenumfangs auf 4.301 erforderlich. Bei Viertelung der Effektgröße und Änderung des Schätzers für die Interventionsgruppe auf 9% ergäbe sich für die Studiengröße ein mehr als 18facher Anstieg auf 18.066 Teilnehmer je Gruppe. Kleine Veränderungen der Effektgröße ziehen große Veränderungen der Studiengröße nach sich.

Häufig lassen sich Untersucher durch die Notwendigkeit riesiger Stichproben bei geringen Ereignisraten abschrecken. Diese Frustration ist zum Teil auf das mangelnde Verständnis dafür zurückzuführen, dass es bei binären Endpunkten eher die Zählerereignisse sind, die die Power einer Studie steigen lassen, als die Werte im Nenner. Nehmen wir als Beispiel $\alpha = 0{,}05$ und eine gewünschte Reduktion der Outcome-Ereignisrate um 40%. Eine Studie mit 2.000 Teilnehmern (1.000 in der Behandlungs- und 1.000 in der Kontrollgruppe) und einer Ereignisrate in der Kontrollgruppe von 10% würde zu einer ähnlichen Trennschärfe führen wie eine Studie mit 20.000 Probanden (10.000 pro Gruppe) bei einer Ereignisrate in der Kontrollgruppe von 1%. Für beide Studien wäre eine vergleichbare Anzahl von Zählerereignissen, nämlich etwa 160, erforderlich, um eine Power von ca. 90% zu gewährleisten.

## Geringe Trennschärfe bei niedrigen Teilnehmerzahlen

Was passiert, wenn die Software bei der Stichprobenberechnung – ausgehend von den sorgfältigen Schätzungen des Untersuchers – eine Studiengröße errechnet, die die Anzahl der verfügbaren Teilnehmer übersteigt? Häufig rechnen Untersucher dann rückwärts und schätzen, dass die Power gemessen an den verfügbaren Teilnehmern zu gering ist (z. B. 0,40) – ein Vorgehen, das wohl eher die Regel als die Ausnahme darstellt [9].

Manche Methodiker raten Ärzten, auf die Durchführung von Studien mit derart niedriger Power zu verzichten. Viele Ethikkommissionen halten Studien mit geringer Trennschärfe für unethisch [10–12]. Zu dieser Auffassung hat auch Chalmers' frühe Arbeit über die mangelhafte Power veröffentlichter Studien beigetragen, was uns wieder an unseren Ausgangspunkt zurückführt.

Chalmers hatte das Gefühl, dass diese Überreaktion durch den Artikel seiner Arbeitsgruppe noch angeheizt wurde [4]. Daraufhin erklärte er schließlich, dass auch Studien mit zu geringer statistischer Power akzeptabel sein können, da sie letztlich in einer Meta-Analyse zusammengefasst werden könnten [4, 13]. Viele Statistiker scheinen sich diesem Standpunkt nicht anzuschließen, und überraschenderweise gilt das auch für diejenigen, die sonst die Durchführung kleiner Studien befürworten [9]. Trotzdem stimmen wir Chalmers' Auffassung zu, auch wenn uns dies mit Sicherheit den Zorn vieler Statistiker und Ethiker eintragen wird.

Unsere Zustimmung erfolgt allerdings unter Vorbehalt:

Erstens sollte eine Studie methodisch fundiert sein und damit systematische Fehler ausschließen. Leider überdeckt diese Beschwörung der adäquaten Power die Diskussion über andere methodische Fragen. So führt beispielsweise auch ein unzureichendes Randomisierungsverfahren zu verzerrten Studienergebnissen, die sich auch dann nicht retten lassen, wenn eine riesige Stichprobe hohe Genauigkeit verspricht [14–16]. Solide angelegte und durchgeführte Studien dagegen liefern selbst bei geringer Power (und Genauigkeit) im Wesentlichen unverzerrte Effektschätzer. Da es sich um unverzerrte Ergebnisse handelt, ließe sich eine solche Studie außerdem mit ähnlichen biasfreien Studien in einer Meta-Analyse zusammenfassen. Und tatsächlich ist dieser Gedanke, vor allem bezüglich prospektiver Meta-Analysen [17], dem Konzept multizentrischer Studien nicht unähnlich.

Zweitens müssen die Autoren ihre Methoden und Ergebnisse angemessen darlegen, um Fehlinterpretationen zu vermeiden. Wenn die Studienergebnisse zweckmäßig mit Hilfe von Intervallschätzern angegeben werden, würde sich eine geringe Trennschärfe korrekt in breiten Konfidenzintervallen um den geschätzten Therapieeffekt niederschlagen. Die Angabe von Konfidenzintervallen liefert wertvolle Information und umgeht ein Problem, das sich aus der allzu einfachen Schlussfolgerung von $p > 0{,}05$ ergibt, nämlich dass „das Fehlen von Evidenz nicht unbedingt beweist, dass es diese Evidenz nicht gibt" [18–20].

Drittens müssen Studien mit geringer Power unabhängig von ihren Ergebnissen publiziert werden, damit sie in Meta-Analysen einfließen können. Publikationsbias gilt als das stärkste Argument gegen Studien mit zu geringer

Trennschärfe [21, 22]. Ein Publikationsbias entsteht, wenn nicht alle durchgeführten Studien auch publiziert werden, weil im Allgemeinen Arbeiten mit statistisch signifikanten Ergebnissen häufiger zur Publikation eingereicht und veröffentlicht werden als Arbeiten mit unklaren Ergebnissen. Studien mit geringer Power verschärfen dieses Problem, weil sie eher zu unklaren Resultaten führen. Sämtliche Studien mit zu geringer Trennschärfe zu verdammen und ihre Durchführung zu verhindern, hieße jedoch relevante Forschung vereiteln. Stattdessen müssen wir gegen die wahren Ursachen des Publikationsbias angehen, und diesbezüglich hat die wissenschaftliche Community auch schon große Fortschritte erzielt. Die Ergebnisse abgeschlossener Studien nicht zu veröffentlichen, gilt in der wissenschaftlichen Literatur mittlerweile sowohl als unwissenschaftlich als auch unethisch [23–25]. Die öffentliche Registrierung von Studien dient der Katalogisierung aller laufenden Studien, damit ihre Ergebnisse nicht verloren gehen. Darüber hinaus wird in verschiedenen Einrichtungen, die sich der Erstellung systematischer Reviews verschrieben haben – hier ist vor allem die Cochrane Collaboration hervorzuheben – nach unveröffentlichten Arbeiten gefahndet, um so dem Publikationsbias entgegen zu wirken.

Erklärungen, die Studien mit zu geringer Power als unethisch brandmarken, kommen uns aus mindestens zwei Gründen ein wenig seltsam vor. Erstens überdeckt die Konzentration auf die Fallzahlen die viel relevanteren Sorgen um die Ausschaltung systematischer Fehler. Und zweitens: Wie kann ein Verfahren, das selbst nur so vor Subjektivität strotzt, hinsichtlich der Studienethik Zündstoff für eine solche Schwarz-Weiß-Entscheidung liefern? Die Studienethik auf die statistische Trennschärfe zu gründen, mutet angesichts dieser Subjektivität allzu simplizistisch und deplatziert an. Und da Untersucher ihre Stichprobengrößen nun einmal auf der Grundlage grober Schätzungen berechnen, werden wir, wenn wir die Durchführung von Studien mit geringer Power als unethisch erachten und diesen Gedanken ins logische Extrem treiben, bald keine Studien mehr durchführen, da die Festlegung des Stichprobenumfangs immer in Zweifel gezogen werden könnte. „Erklärungen, dass die Durchführung kontrollierter Studien unethisch sei, wenn nicht ein willkürlich festgelegtes Maß an statistischer Power gewährleistet ist, ergeben keinen Sinn, wenn die Alternative heißt, Unwissenheit hinsichtlich der Wirkungen medizinischer Interventionen in Kauf zu nehmen" [24]. Aussagen, dass Studien mit zu geringer Power unethisch sind, stellen die Vernunft in Frage und lassen ferner außer Acht, dass potenzielle Probanden gelegentlich auch an Studien teilzunehmen wünschen [26].

## Der „Tanz" um die Stichprobengröße

Manchmal veranstalten Untersucher, um die richtige statistische Trennschärfe zu erzielen, um die Stichprobengröße geradezu einen „Eiertanz" [27, 28]. Dazu gehört auch, dass die Schätzer der verschiedenen Parameter (vor allem für den Therapieeffekt, den es nachzuweisen lohnt) nachträglich an die zur Verfügung stehenden Studienteilnehmer angepasst werden. In der Praxis ist dieses Vorgehen, wie wir selbst und auch andere erfahren haben, recht häufig anzutreffen [27]. Möglicherweise unterstützen auch Kostenträger, Protokoll- und sogar Ethikkommissionen ein solch retrospektiv orientiertes Vorgehen. In der Tat handelt es sich dabei um einen funktionierenden Lösungsansatz für ein echtes Problem. Angesichts der Umstände wollen wir mit diesem „Eiertanz" hier nicht allzu hart ins Gericht gehen, da er wahrscheinlich die Durchführung vieler wichtiger Studien erleichtert hat. Außerdem zeichnet dieses Vorgehen, wenn die oben beschriebenen Werte bekannt sind, ein wahres Bild vom erforderlichen Stichprobenumfang. Dennoch unterstreicht ein solches Vorgehen, wie widersprüchlich das Argument ist, Studien mit geringer Trennschärfe seien unethisch: Eine geplante Studie gilt vor dem „Eiertanz" als unethisch und genügt plötzlich den ethischen Ansprüchen, nur weil der Schätzer der Effektgröße verändert wurde. Schließlich lassen sich für alle Studien beliebig viele Trennschärfen angeben, und der Begriff „geringe Trennschärfe" ist relativ.

## Änderung der Stichprobengröße

Stehen zusätzliche Studienteilnehmer und flexible Ressourcen zur Verfügung, könnte der Untersucher eine Strategie zur Änderung seines Stichprobenumfangs in Erwägung ziehen, mit der sich einige der durch die groben Schätzungen bei der ursprünglichen Berechnung der Stichprobengröße bedingten Probleme verringern ließen. Gewöhnlich führen solche Modifikationen zur Zunahme der Stichprobengröße [29], sodass dem Untersucher weitere Probanden und Finanzmittel zur Verfügung stehen sollten, um eine entsprechende Anpassung vornehmen zu können.

Verschiedene Ansätze zur Durchführung solcher Modifikationen stützen sich auf die Korrektur der Ereignisrate, die Veränderung des Endpunktes oder des Behandlungseffekts [30–33]. Wichtig ist, dass zur Vermeidung von Bias jede Änderung des Stichprobenumfangs in einem Zwischenstadium der Studie nach einem zuvor festgelegten Plan ablaufen sollte. Zu empfehlen ist, dass von den Trägern der Studie oder der Studienleitung bezüglich Zeitpunkt und Methodik eventueller Modifikationen im Studienprotokoll ein nachvollziehbarer Plan aufgestellt wird [31].

## Die Sinnlosigkeit der nachträglichen Berechnung der Power

Eine Studie ergibt einen Behandlungseffekt und ein Konfidenzintervall für die Ergebnisse, in dem die Power der Studie zum Ausdruck kommt. Damit muss uns die Power nicht mehr weiter kümmern [7, 27, 34]. Trotzdem stellen manche Untersucher nach Beendigung einer Studie mit statistisch nicht signifikanten Ergebnissen auf der Grundlage der vorgefundenen Werte für die ge-

schätzten Parameter noch einmal Berechnungen zur Studienpower an. Diese Übung hat zwar ihren besonderen Reiz, ist jedoch überflüssig, zumal das Ergebnis, nämlich „geringe Power", tautologisch wäre [7, 27]. Mit anderen Worten, wer so vorgeht, ist schlecht beraten, denn er erhält lediglich Antwort auf eine bereits beantwortete Frage.

## Worauf sollte der Leser bei der Fallzahlschätzung achten?

Der Leser sollte nach der *a priori* geschätzten Stichprobengröße suchen. Aufschluss über die Power der Studie geben uns in Studienberichten die Konfidenzintervalle. Trotzdem liefern auch Stichprobenberechnungen relevante Informationen. Erstens spezifizieren sie nämlich den primären Endpunkt, was uns davor schützt, dass Endpunkte nachträglich geändert werden und für einen gar nicht als primären Endpunkt geplanten Zielparameter plötzlich ein großer Effekt behauptet wird [35]. Zweitens macht das Wissen um die geplante Studiengröße den Leser auf potenzielle Probleme aufmerksam. Hatte die Studie mit Rekrutierungsproblemen zu kämpfen? Wurde die Studie wegen eines statistisch signifikanten Ergebnisses vorzeitig abgebrochen? Wenn ja, sollten die Autoren eine formale statistische Stoppregel angegeben haben [36]. Benutzen die Autoren nämlich keine formale Stoppregel, dann führt die Mehrfachbearbeitung der Studiendaten zu einer Aufblähung von $\alpha$ [5, 29]. Zu ähnlichen Problemen kann es bei Stichproben kommen, die größer als geplant ausgefallen sind. Die Angabe geplanter Größen, so willkürlich sie auch sein mag, legt das Fundament für eine transparente Studienberichterstattung.
Die Angabe einer niedrigen Studienpower oder fehlende Angaben zur Fallzahlschätzung stellen im Allgemeinen keinen verhängnisvollen Fehler dar. Eine geringe Power kann einen Mangel an methodischem Wissen widerspiegeln, ebenso gut aber auch bloß Hinweis auf eine unzureichende Anzahl potenzieller Studienteilnehmer sein. Stichprobenberechnungen enthalten selbst bei geringer Power immer noch, wie oben beschrieben, essenzielle Informationen. Was aber, wenn die Autoren es versäumen, ihre *a priori* durchgeführten Stichprobenberechnungen zu erwähnen? Wenn keine Angaben zu primären Endpunkten und Stoppregeln gemacht werden, sollte der Leser die Ergebnisse solcher Studien mit Vorsicht interpretieren. Überdies lässt das Versäumnis, über die Berechnung der Stichprobengröße zu berichten, auf eine methodische Naivität schließen, die ganz andere Probleme erwarten lässt.
In jedem Fall aber sollte der Leser unbedingt auf die von den Untersuchern verborgenen systematischen Fehler (Bias) achten. Autoren, die nicht auf eine mangelhafte Randomisierung, eine unzureichende Verdeckung der Behandlungszuteilung, ungenügende Verblindung oder die mangelhafte Bereitschaft der Teilnehmer, in der Studie zu verbleiben, aufmerksam machen, haben Unzulänglichkeiten zu verbergen, die schwerwiegende systematische Fehler nach sich ziehen können [37–41]. Aus zwei wichtigen Gründen sollte sich der Leser, wenn er auf ungenügende Stichprobengrößen stößt, deshalb auch nicht allzu sehr sorgen. Erstens führen diese nicht zu systematischen Fehlern, und zweitens kommt jeder Zufallsfehler in den Konfidenzintervallen und *p*-Werten ans Tageslicht. Die größten Probleme verursachen verborgene systematische Fehler. Mit anderen Worten: Der Leser sollte eine Studie nicht einfach wegen ihrer geringen statistischen Power außer Acht lassen, sondern ihren Wert entsprechend sorgfältig abwägen. Und dieser Wert ergibt sich im Kontext anderer (früherer oder zukünftiger) Forschungsarbeiten [42].
Ratsam ist, nach allen Annahmen zu suchen, die der Berechnung des Stichprobenumfangs zugrunde liegen, also Typ I-Fehler ($\alpha$), Power (oder $\beta$), Ereignisrate in der Kontrollgruppe und interessierender Therapieeffekt (oder analog eine Ereignisrate in der Behandlungsgruppe). Eine Aussage wie „Die Berechnung der erforderlichen Fallzahl von 120 pro Gruppe erfolgte mit $\alpha = 0{,}05$ und einer Power von 0,90" ist nahezu bedeutungslos, da sie die Schätzer der Effektgröße und der Ereignisrate in der Kontrollgruppe unberücksichtigt lässt. Selbst in kleinen Studien kann die Power hoch genug sein, um große Therapieeffekte nachweisen zu können.
Ferner sollte der Leser die bei der Berechnung der Stichproben getroffenen Grundannahmen unter die Lupe nehmen. Denn vielleicht ist in seinen Augen eine kleinere Effektgröße angemessener als die von den Autoren geplante Effektgröße. Damit wäre sich der Leser auch der im Verhältnis zu der von ihm bevorzugten Effektgröße geringeren Studienpower bewusst.

## Fazit

Power oder Trennschärfe ist ein wichtiges statistisches Konzept, das jedoch von seinem überzogenen ethischen Stellenwert befreit werden sollte. Wir wenden uns dagegen, dass Studien allein aufgrund einer inhärent subjektiven, ungenauen Fallzahlschätzung als unethisch gebrandmarkt werden. Wir befürworten, Studien so zu planen, dass eine adäquate Trennschärfe erzielt werden kann, und begrüßen große multizentrische Studien von der Art einer ISIS-2 [43]; tatsächlich sollten viel mehr solcher Studien durchgeführt werden. Würde die akademische Welt jedoch ausschließlich auf große Studien pochen, blieben viele unbeantwortete Fragen in der Medizin auch weiterhin ohne Antwort. Eine Verlagerung des Schwerpunktes von der Fixierung auf den Stichprobenumfang hin zur methodischen Qualität würde uns mehr Studien mit weniger Bias bescheren. Unverzerrte Studien mit unpräzisen Ergebnissen sind immer noch besser als gar keine Ergebnisse.

### Offenlegung von Interessenkonflikten

Hiermit erklären wir, dass keinerlei Interessenkonflikte vorliegen.


### Danksagung

Wir bedanken uns bei David L. Sackett, Douglas G. Altman, Willard Cates und Sir Iain Chalmers für ihre hilfreichen Anmerkungen zu einer früheren Version des vorliegenden Manuskriptes.

## Literatur


[1] Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group trials. Lancet 2001;357:1191–4.

[2] Altman DG, Schulz KF, Moher D et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001; 134:663–94.

[3] Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. N Engl J Med 1978; 299:690–4.

[4] Sackett DL, Cook DI. Can we learn anything from small trials? Ann NY Acad Sci 1993;703:25–31.

[5] Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.

[6] Meinert CL, Clinical trials: design, conduct, and analysis. New York: Oxford University Press, 1986.

[7] Piantadosi S. Clinical trials: a methodologic perspective. New York: John Wiley and Sons, 1997.

[8] Sinei SK, Schulz KF, Lamptey PR et al. Preventing IUCD-related pelvic infection: the efficacy of prophylactic doxycycline at insertion. Br J Obstet Gynaecol 1990;97:412–9.

[9] Matthews JN. Small clinical trials: are they all bad? Stat Med 1995;14: 115–26.

[10] Edwards SI, Lilford RJ, Braunholtz D, Jackson J. Why "underpowered" trials are not necessarily unethical. Lancet 1997;350:804–7.

[11] Halpern SD, Karlawish JH. Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358–62.

[12] Lilford RJ. The ethics of underpowered clinical trials. JAMA 2002;288:2118–9.

[13] Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline, I: control of bias and comparison with large co-operative trials. Stat Med 1987;6:315–28.

[14] Peto R. Failure of randomisation by 'sealed' envelope. Lancet 1999;354: 73.

[15] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273:408–12.

[16] Schulz KF. Subverting randomization in controlled trials. JAMA 1995;274: 1456–8.

[17] Walker MD. Atrial fibrillation and antithrombotic prophylaxis: a prospective meta-analysis. Lancet 1989;1:325–6.

[18] Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.

[19] Detsky AS, Sackett DL. When was a "negative" clinical trial big enough? How many patients you needed depends on what you found. Arch Intern Med 1985;145:709–712.

[20] Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases: a way out of a conundrum. BMJ 1995;311: 1621–5.

[21] Dickersin K. How important is publication bias? A synthesis of available data. AIDS Educ Prev 1997;9:15–21.

[22] Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr. Publication bias and clinical trials. Control Clin Trials 1987;8:343–53.

[23] Chalmers I. Underreporting research is scientific misconduct. JAMA 1990;263: 1405–8.

[24] Chalmers I. Cardiotocography v Doppler auscultation: all unbiased comparative studies should be published. BMJ 2002;324:483–5.

[25] Antes G, Chalmers I. Under-reporting of clinical trials is unethical. Lancet 2003;361:978–9.

[26] Chalmers I. What do I want from health research and researchers when I am a patient? BMJ 1995;310:1315–8.

[27] Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994;121:200–6.

[28] Peipert JF, Metheny WP, Schulz K. Sample size and statistical power in reproductive research. Obstet Gynecol 1995; 86:302–5.

[29] Ellenberg SS, Fleming TR, DeMets DL. Data monitoring committees in clinical trials. Chichester: John Wiley and Sons, 2002.

[30] Wang SJ, Hung HM, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. Stat Med 2001;20:1903–12.

[31] Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. Biometrics 1999;55: 853–7.

[32] Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics 1999;55: 1286–90.

[33] Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Stat Med 1990;9: 65–71.

[34] Fayers PM, Machin D. Sample size: how many patients are necessary? Br J Cancer 1995;72:1–9.

[35] Chan AW, Hrobjartsson A, Haahr MT, Goetzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291:2457–65.

[36] Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet (in press).

[37] Schulz KF, Grimes D. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet 2002;359:781–5.

[38] Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002;359:696–700.

[39] Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002; 359:614–8.

[40] Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002; 359:515–9.

[41] Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. Lancet 2002;359: 966–70.

[42] Clarke M, Alderson P, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals. JAMA 2002;287: 2799–801.

[43] ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase: oral aspirin, both, or neither among 17.187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 1988;2:349–60.



**Korrespondenzadresse:**
Dr. Kenneth F Schulz
KSchulz@fhi.org


**Anmerkung der Redaktion:**
Die Übersetzung dieses Artikels erfolgte durch Frau Karin Beifuss (Stuttgart), die fachliche Bearbeitung übernahm Frau Gerta Rücker (IMBI – Institut für Medizinische Biometrie und Statistik Universitätsklinikum Freiburg). Beiden sei an dieser Stelle sehr herzlich gedankt.