Methodenlehre (13) Power-Analyse: die Magie der Statistik – Oder: Der Unterschied zwischen Signifikanz und Relevanz

Harald Walach (http://harald-walach.de/methodenlehre-fuer-anfaenger/) - Mai 2014

Normalerweise ist der Durchschnittsbürger und Durchschnittswissenschaftler zufrieden, wenn er hört, ein Forschungsergebnis sei „statistisch signifikant" gewesen. Wir meinen dann landläufig: die Hypothese, mit der man an die Forschung ging, sei belegt, das Faktum, das man untersucht bewiesen. Und umgekehrt, findet man kein signifikantes Ergebnis, glaubt man, das in Frage stehende Phänomen sei nicht gefunden, also nicht existent. Deswegen glaubt z.B. der Durchschnittsarzt, -journalist und -bürger die Bioresonanz sei als unwirksam belegt und Homöopathie ist Placebo, und halb Amerika nimmt Lipidsenker zur Primärprävention von Herzinfarkt, weil man glaubt das sei eine wissenschaftlich bewiesene Tatsache. Ich will in diesem Blog ein paar von diesen Meinungen genauer unter die Lupe nehmen und zeigen, warum sie entstanden sind und die Frage stellen, wie berechtigt sie sind. Es wird sich herausstellen: es hat mit dem zu tun, was ich die Magie der Statistik nenne. Das ist die Frage, wie mächtig ein statistischer Test ist. Die hängt zusammen mit der Frage, wie groß der Effekt ist, den wir untersuchen. Und davon hängt ab, wie groß die Stichprobe ist, die wir benötigen, um den Effekt wirklich statistisch sichtbar zu machen, oder ein signifikantes Ergebnis zu erhalten. Anders ausgedrückt: *Wenn* es einen systematischen Effekt gibt, egal wie groß er ist, dann lässt er sich mit einer Untersuchung auch belegen, vorausgesetzt, wir haben genügend Ressourcen. Die Frage, die sich jeder Leser einer wissenschaftlichen Untersuchung stellen sollte ist nicht: Ist eine Studie signifikant? Sondern: Ist der gezeigte Effekt, egal ob signifikant oder nicht, klinisch und systematisch von Bedeutung? Wenn er dann auch noch signifikant ist, können wir von einer wissenschaftlichen Bestätigung ausgehen. Wenn er nicht signifikant ist, müssen wir uns die Frage stellen: war die Größe der Studie geeignet, den Effekt zu finden? bzw. umgekehrt: wie groß müsste eine Studie sein, um einen Effekt von der gefundenen Größenordnung mit einigermaßen zufriedenstellender Sicherheit statistisch absichern zu können? Das ist die Essenz der Power-Analyse, um die es jetzt geht.

Wir haben es also in jeder wissenschaftlichen Untersuchung mit dem Spiel von insgesamt vier Größen zu tun, die voneinander abhängen wie die Teile eines filigranen Mobile. Wenn wir eines verändern, verändern sich alle anderen auch. Das wären:

1. Der Fehler erster Art oder der alpha-Fehler.
2. Der Fehler zweiter Art oder der beta-Fehler.
3. Die Größe des Effekts, oder die Effektgröße.
4. Die Größe der Studie oder die Anzahl von untersuchten Personen (im Falle von klinischen oder diagnostischen Studien) bzw. die Anzahl der Beobachtungen.

1. *Der Alpha-Fehler*:

Inhaltlich meint er den Fehler, den wir machen, wenn wir behaupten es gäbe irgendwo einen Effekt, obwohl er nicht da ist. Wenn wir also beispielsweise behaupten, Homöopathie oder Bioresonanztherapie sei wirksam und tatsächlich sind sie es nicht, machen wir einen Fehler erster Art oder einen Alpha-Fehler. Die Konvention der Forschung geht in der Regel davon aus, dass wir einen solchen Fehler nur in 5 von 100 Fällen bereit sind zu akzeptieren. Das ist der Grund, weswegen viele Studien das Signifikanzniveau auf alpha = 0.05 setzen. Wenn also eine Studie findet $p < 0.05$ und der Wissenschaftler folgert, das Ergebnis sei signifikant, dann meint dies, in Worten: Wenn wir aufgrund

der Studie behaupten, die untersuchte Intervention sei wirksam, oder der gefundene Effekt vorhanden, oder die untersuchte Diagnostik trennscharf, dann machen wir in 5 von 100 Fällen einen Fehler. Deswegen sagen wir auch: die Irrtumswahrscheinlichkeit beträgt 5%. Wir können auch sagen: Mit einer 5%igen Wahrscheinlichkeit kann ein solches Ergebnis per Zufall zustande kommen. Deswegen werden wir als Wissenschaftler in besonders wichtigen Fällen auch darauf bestehen, das Alpha-Niveau herauf zu setzen, sagen wir auf 1%. Dann machen wir einen solchen Fehler nur in einem von 100 Fällen. Entsprechend ist ein Effekt, der auf einem Signifikanzniveau von $p < 0.0001$ sichtbar wird mit einem sehr kleinen Alpha-Fehler behaftet bzw. die Wahrscheinlichkeit, dass wir einen Fehler machen, wenn wir behaupten ein solcher Effekt existiert, ist nicht sehr groß. Oder: ein solcher Effekt tritt per Zufall nur sehr selten auf.
Die Wissenschaft ist, als Ausdruck kollektiven Bemühens der Gesellschaft, daran interessiert sich davor zu hüten, falsche Behauptungen aufzustellen. Das würde sie tun, wenn sie einen Effekt behaupten würde, wo gar keiner vorhanden ist. Daher legen alle Wissenschaftler, Herausgeber von Zeitschriften und die öffentliche Meinung großen Wert auf die Kontrolle des alpha-Fehlers. Denn immerhin würde man dann ja eine möglicherweise gefährliche, kostspielige oder anderweitig aufwändige Prozedur befürworten, obwohl es dafür keinen sachlichen Grund gibt. Deswegen wissen auch fast alle Leute über die Bedeutung der statistischen „Signifikanz“ Bescheid, oder anders ausgedrückt, die Bedeutung des alpha-Fehlers ist allen einigermaßen klar.
Dies hängt auch mit der Logik statistischen Testens zusammen, die u.a. auf Fisher zurückgeht [1]: Hier wird eine Hypothese formuliert und zwar der Form: Behandlung x ist nicht verschieden von Kontrollbehandlung y. Homöopathie ist gleich wirksam wie Placebo etwa. Das ist die sog. „Null-Hypothese“, also die Vermutung, dass kein Unterschied oder kein Effekt vorhanden ist. Nun kann man über die statistische Prozedur des Testens einer solchen Hypothese herausfinden, ob das vermutlich richtig ist oder falsch. Dazu stellt man eine Alternativ-Hypothese auf, also z.B. „Homöopathie ist besser als Placebo“ und untersucht nun den empirisch gefundenen Unterschied daraufhin, ob er mit der ursprünglichen Null-Hypothese, also dass kein Unterschied vorhanden ist, kompatibel ist. Die Prozedur dazu ist ein statistischer Test. Was dahinter genau steckt, will ich in einem anderen Blog besprechen. Nehmen wir der Einfachheit halber an, unsere Untersuchung hätte ergeben, dass die Wahrscheinlichkeit, dass die Alternativ-Hypothese stimmt, dass es also einen Unterschied zwischen Homöopathie und Placebo gibt $p > 0.05$ ist, also wir in mehr als 5 von 100 Fällen einen Fehler machen, wenn wir einen solchen Unterschied behaupten. Dann weist der statistische Test unsere Alternativ-Hypothese zurück und sagt uns konservativerweise, wir sollten besser die Null-Hypothese beibehalten und davon ausgehen, dass es keinen Unterschied zwischen sagen wir, Bioresonanz-Therapie und Placebo gibt. Dann sagt man: es gibt keinen wissenschaftlichen Hinweis auf einen Unterschied zwischen Bioresonanz und Placebo. Heißt das nun, es gibt ihn nicht? Nicht unbedingt, denn wir müssen noch die drei anderen Größen in Rechnung stellen. Denn es könnte sein, dass wir einen Fehler der zweiten Art begehen, einen Beta-Fehler.

2. *Der Beta-Fehler*:

Dieser besteht darin, dass wir einen Effekt, der vorhanden ist, übersehen, dass wir also fälschlicherweise behaupten, ein Effekt wäre nicht vorhanden. Wenn wir also sagen würden, Bioresonanz-Therapie ist eine Placebo-Therapie, obwohl diese Behauptung in Wirklichkeit falsch ist, dann würden wir einen Fehler der zweiten Art oder einen Beta-Fehler machen. Während der Alpha-Fehler die konservative Seite der Wissenschaft abbildet, also die Sorge keine falsche Behauptungen aufzustellen, spiegelt der Beta-Fehler die mangelnde Sensitivität einer Untersuchung wider, also die mangelnde Sorgfalt beim Hinsehen, das Übersehen von Effekten, weil die angewandten Instrumente nicht fein genug auflösen. Z.B. war die Tatsache, dass man lange nicht wusste, dass Infektionen von

Bakterien ausgelöst wurden, technisch gesprochen ein Beta-Fehler, den man deswegen gemacht hat, weil man keine hochauflösenden Mikroskope hatte, genauso wie die Unwissenheit über Viren als mögliche Krankheitsursachen vor der Erfindung des Elektronenmikroskopes ein Beta-Fehler war, weil man sie nicht sehen oder nachweisen konnte. Der Beta-Fehler hängt also damit zusammen, dass Effekte übersehen werden. Der Punkt ist nun folgender, wie man intuitiv leicht nachvollziehen kann: Den Alpha-Fehler können wir willkürlich festsetzen. Der Beta-Fehler ist in inniger Weise mit der Größe des zu untersuchenden Effekts verbunden. Wir können noch so gern sagen: wir wollen keine Krankheitsursache übersehen. Wenn wir kein Elektronenmikroskop oder kein immunologisches Assay haben können wir die winzigen Viren nicht sehen und behaupten, es gäbe keinen Grund für eine Erkrankung. Wir machen dann einen Beta-Fehler, wenn in Wirklichkeit Viren Krankheitsauslöser sind, die wir aber aufgrund fehlender Instrumente nicht entdecken können. Ähnlich der Beta-Fehler in einer klinischen Studie: wir übersehen gern Effekte, die kleiner sind als erwartet, bzw. kleiner, als mit der vorhandenen Studiengröße sichtbar gemacht werden kann. Sind sie deswegen irrelevant? Das Beispiel mit den Bakterien, die man mit dem Lichtmikroskop sehen kann und mit den Viren, die man nur mit dem Elektronenmikroskop bzw. einem immunologischen Assay sehen kann zeigt, dass das offenbar falsch ist. Denn Viren können genauso bedeutsam sein wie Bakterien und Bakterien können genauso gefährliche Krankheiten auslösen wie eine Verletzung mit einem Messer, das man sehen kann.
Der Beta-Fehler hängt also innig mit der Größe des zu erwartenden oder zu untersuchenden Effektes zusammen. Wir können allenfalls sagen: wir wollen einen Effekt, falls er denn vorhanden ist, allenfalls mit einer Wahrscheinlichkeit von ebenfalls 5% oder 10% übersehen, also in 5 oder 10 von Hundert Fällen einen Fehler machen, wenn wir behaupten, den Effekt gäbe es gar nicht, also etwa Bioresonanz sei unwirksam. Aber können wir das in der gleichen Weise formalisieren? Ja und nein. Nein insofern, als es nicht unabhängig von anderen Größen geht. Ja insofern, dass wir eben die zu erwartende Effektgröße in unsere Rechnung mit einkalkulieren müssen. Deswegen müssen wir uns eben Rechenschaft geben über die zu untersuchende Effektgröße oder Effektstärke (ES).

3. *Die Effektgröße*

Effektgrößen sind numerische Maße die angeben, wie stark sich zwei Gruppen, z.B. eine behandelte und eine unbehandelte, oder eine mit richtiger und eine mit Scheinarznei behandelte Gruppe unterscheiden (im Prinzip kann man auch einen Zusammenhang zwischen Variablen, also einen Korrelationskoeffizienten als Effektgröße angeben; da das in der klinischen Forschung eher unüblich ist, gehe ich jetzt nicht darauf ein). Man spricht dann von einer Effektgröße, die den Unterschied *zwischen* zwei Gruppen angibt, im Englischen als „between-group effect size (ES)“ bezeichnet. Man kann auch die Größe des Effekts quantifizieren, den man sieht, wenn man nur eine Gruppe zu zwei verschiedenen Meßzeitpunkten untersucht, zwischen denen irgend etwas passiert ist, z.B. eine natürliche Veränderung, wenn es um Reifung und Wachstum geht, oder eine Intervention. Das ist dann eine Effektgröße *innerhalb* einer Gruppe, die von vorher zu nachher, oder von prä zu post festgestellt wird und im Englischen mit „within-group ES“ bezeichnet wird. Eine solche ES innerhalb ein und derselben Gruppe ist zwar numerisch ähnlich zu ermitteln hat aber systematisch klarer Weise eine andere Bedeutung, weil sie nämlich keinen Unterschied zwischen Gruppen quantifiziert, sondern innerhalb ein und derselben Gruppe. Deswegen ist sie für die weiteren Überlegungen hier nicht mehr von Bedeutung, und wann immer ich im folgenden von „Effektgröße“ rede, werde ich erstens öfterdie Abkürzung ES für „effect size“ verwenden und zweitens die ES *zwischen* zwei Gruppen damit meinen.

Zwei Familien von Effektgrößen:

Was ebenfalls wichtig ist, zwar nicht prinzipiell, aber für das konkrete Verständnis, ist die Tatsache, dass es zwei grundsätzlich verschiedene Effektgrößen gibt. Dies hängt damit zusammen, dass man Ereignisse in dieser Welt sehr grob in zwei Kategorien einteilen kann: solche, deren Vorkommen oder Nichtvorkommen man feststellen kann, und solche, die man genauer messen kann. Ereignisse, die entweder vorkommen oder nicht, nennen wir auch dichotom. Denn sie lassen sich mit einer ja/nein Entscheidung quantifizieren: Entweder es regnet, oder es regnet nicht. Entweder jemand lebt oder er ist tot. Entweder ist jemand geheilt oder immer noch krank, hat ein bestimmtes klinisch definiertes Ergebnis erreicht oder nicht. Solche Merkmale werden mit Variablen erfasst die man mit 1 oder 0 kodieren kann, anstelle von „ja“ oder „nein“. Deswegen heißen sie eben dichotome Variable, ein spezieller Fall einer sogenannten nominalen Skalierung, also einer Darstellung, bei der Zahlenwerten keine numerische, sondern eine besonders definierte inhaltliche Bedeutung zukommt. In diesem Fall eben „1“ für „Ereignis festgestellt“, was auch immer das Ereignis ist, und „0“ für „Ereignis nicht festgestellt“. Solche Ergebnisdarstellungen sind in der Medizin relativ häufig, etwa in der Krebsforschung, wenn es um Überleben geht, oder wenn man ein robustes Ergebnismaß will, das dem behandelnden Arzt Klarheit darüber geben soll, wie viel Prozent einer Gruppe von Patienten „geheilt“ sind im Sinne einer bestimmten Definition. Effektgrößen dieser Familie werden grundsätzlich durch Verhältnisse ausgedrückt. So etwa durch das Verhältnis von Ereignissen in einer Gruppe im Verhältnis zu dem Nichtvorkommen solcher Ereignisse in dieser Gruppe und dieses Verhältnis wiederum im Verhältnis zu den gleichen Proportionen in der Vergleichsgruppe, numerisch also etwa RR = a/n1 : a'/n2. RR ist die sog. rate ratio. Sie gibt an, wie das Verhältnis von gebesserten (a) zur Gesamtzahl in der Behandlungsgruppe zum gleichen Verhältnis in der Kontrollgruppe (a'; n2) ist. Ist RR = 2, dann bedeutet das: in der Behandlungsgruppe wurden zweimal so viel Leute gesund wie in der Kontrollgruppe; analog bedeutet eine HR (hazard ratio) von 0.5, dass in der behandelten Gruppe nur die Hälfte aller unerwünschten Ereignisse, z.B. Tod oder Herzinfarkt, aufgetreten ist [2]. Etwas anders definierte Masse dieser Familie sind die Odds-Ratio (OR), oder die logarithmierte Odds Ragio (log-OR). Ich gehe auf Effektgrößenbestimmung in einem anderen Blog noch gesondert ein.

Eine grundsätzlich andere Familie von Effektgrößen stammt aus Variablen, die kontinuierlich gemessen werden, sog. „intervallskalierten“ Variablen. Die bekanntesten intervallskalierten Variablen sind die Temperatur, oder die Zeit. Jeder Abstand zwischen den Größen misst ein ähnlich großes Intervall und ordnet die Größenverhältnisse Zahlen zu. In der Medizin oder Psychologie versucht man ebenfalls, möglichst oft solche Messungen vorzunehmen, weil sie intuitiv präziser sind, als dichotome Merkmale. So ist etwa die Aussage: „In Berlin hat es 5°, in Singapore 28° C“ präziser als die Aussage „In Berlin ist es kalt, in Singapore ist es warm“. Typische Beispiele für intervallskalierte Maße sind der Blutdruck, der in mm Quecksilbersäule (Hg) gemessen wird, oder die Herzschläge pro Minute, oder die Zeit, die ein Signal oder eine Reaktion braucht in Millisekunden, oder die Schmerzstärke in mm einer visuellen Analogskala (VAS). Durch Psychometrie, also die Kunst, komplexe Konstrukte wie Lebensqualität, Depression, Angst, usw. in numerische Werte zu übersetzen, können auch solche Inhalte meßbar gemacht werden. Das tun typischerweise psychometrische Skalen, die gut konstruiert und empirisch validiert sind.
Effektgrößen aus Variablen dieser Kategorie, gewöhnlich als „d“ bezeichnet – abgekürzt für „difference“ -, kann man gewinnen, indem man die Mittelwerte zweier Gruppen voneinander subtrahiert. Weil alle möglichen Meßverfahren auf sehr unterschiedlichen Skalen messen – Größe wird etwa in cm oder mm gemessen, Zeit in Sekunden, Jahren, oder Millisekunden, Schmerz oft in mm einer VAS, Lebensqualität, Intelligenz oder Depression in Skalen ohne eigentliche Benennung – muß man sich einen Trick einfallen lassen, um sie vergleichbar zu machen. Dies ist der statistische Trick der „Standardisierung“. Man erreicht dies, indem man die Differenz durch die Standardabweichung der entsprechenden Maße teilt. Weil Mittelwerte und Standardabweichungen

einer bestimmten Meßgröße immer in einem gewissen Verhältnis zueinander stehen, gewinnt man auf diese Weise Vergleichbarkeit.

Die entsprechende grundlegende Formel ist also **d = m1-m2/sd** [3],

wobei m1 der Mittelwert der einen Gruppe,
m2 der Mittelwert der anderen Gruppe ist und
sd die Standardabweichung.

Wir können nun also prinzipiell verstehen, wie die Größe eines Effektes in einer Studie erfasst werden kann. Für unseren Kontext ist es nun wichtig zu verstehen: Wenn wir einen bestimmten Alpha-Fehler definiert haben, in der Regel 5%, dann hängt unsere Fähigkeit, einen Effekt zu entdecken davon ab, wie groß der Effekt ist. Denn je größer er ist, umso leichter sehe ich ihn. Um zu erkennen, dass ein Messer eine Wunde verursacht, muss ich nur die Augen aufmachen. Um zu sehen, wie ein Bakterium eine Zelle infiziert, brauche ich ein Lichtmikroskop. Um zu sehen, wie Viren die Erbsubstanz einer Zelle umfunktionieren brauche ich immunologische Methoden. Wenn ich also einen großen Effekt habe, kann ich mit relativ wenig Personen das Vorhandensein des Effekts so sichtbar machen, dass ein Irrtum ziemlich unwahrscheinlich ist. Ist der Effekt aber nicht ganz so groß, muß ich natürlich eine größere Stichprobe heranziehen. Damit sind wir bei der vierten wichtigen Variablen:

4. *Die Stichprobengröße*

Unsere Fähigkeit, in einer klinischen oder diagnostischen Studie einen therapeutischen Effekt sichtbar zu machen oder eine Trennung zweier Gruppen vorzunehmen hängt also davon ab, wie groß der Effekt ist. Diese Effektgröße definiert, wie groß die Stichprobe sein muß, die ich benötige, um den Effekt sichtbar zu machen. Die Stichprobengröße ist dabei so etwas wie die Feinheit oder Grobheit unseres Instrumentes. Eine kleine Stichprobe ist vergleichbar dem Sehen mit bloßem Auge. Eine relativ große Stichprobe ist, als ob wir ein Hilfsmittel, eine Brille oder ein Mikroskop etwa, benötigen, um etwas zu sehen. Eine sehr große Stichprobe ist wie ein hochauflösendes Hilfsmittel wie ein Elektronenmikroskop, eine immunologische Probe oder ein Radioteleskop, um einen sehr kleinen Effekt sichtbar zu machen. Heißt das dann, dass man sich den Aufwand auch sparen könnte? Diese Frage kann nur klinisch und nicht statistisch beantwortet werden. Die Antwort auf die Frage hängt davon ab, wie wichtig der Effekt ist. Wenn wir eine relativ leicht zu behandelnde Krankheit vor uns haben oder eine, die von selber gut wird, wie etwa Schnupfen oder Grippe, dann wird uns nur ein wirklich großer Effekt interessieren. Wenn wir eine extrem gefährliche oder unbehandeltbare Krankheit vor uns haben, wie einen aggressiven Krebs oder schwere Schmerzen, dann wird auch ein relativ kleiner Effekt von Bedeutung sein.

Die Power oder statistische Mächtigkeit [4]

Die Power eines Tests oder die statistische Mächtigkeit ist nun die Wahrscheinlichkeit, einen Effekt mit einer Studie definierter Größe zu entdecken, wenn er denn vorhanden ist. Sie ist definiert als 1-beta, wobei beta der Beta-Fehler oder Fehler zweiter Art ist. Wenn ich also bereit bin, in 10 von 100 Fällen in Kauf zu nehmen, dass ich einen Effekt, wenn er vorhanden ist, übersehe, dann mache ich potenziell einen beta-Fehler von 0.10. Entsprechend wäre die Power 1-beta = 0.9 oder 90%. Wie wir oben gesehen haben, kann ich aber, anders als beim alpha-Fehler, den beta-Fehler und damit die Power nicht einfach definieren. Vielmehr hängt beides innig damit zusammen, wie groß der Effekt ist,

den ich zeigen will. Dies bedingt nämlich, wie viele Personen ich in eine Studie aufnehmen muss. Habe ich einen großen Effekt, dann ist bei gleichem alpha-Fehler die Anzahl benötigter Studienteilnehmer relativ klein, selbst wenn ich nur wenig Gefahr laufen will, den Effekt zu übersehen und damit eine große statistische Mächtigkeit meines Tests fordere. Ist der Effekt, den ich zu sehen erwarte relativ klein, dann muß ich bei gleichem Alpha-Fehler eine viel größere Anzahl von Personen in meine Studie rekrutieren, um den Beta-Fehler ebenfalls gering zu halten und damit den Effekt mit einer einigermaßen vernünftigen statistischen Mächtigkeit zu sehen. Bildlich gesprochen: je kleiner der Effekt, umso feiner die Instrumente, um ihn zu sehen und je größer der Aufwand, der getrieben werden muß um die Sichtbarkeit gleich zu halten. Es wäre ein Fehler zu behaupten – nämlich ein Beta-Fehler – von Bakterien kann man nicht krank werden, nur weil man sie mit bloßem Auge nicht sehen kann. Es war ebenfalls ein Fehler zu sagen, es gäbe keine Planetensysteme ausserhalb unseres Sonnensystem (eine Behauptung, die mal vor ca. 20 Jahren oft zu hören war), nur weil wir keine Teleskope hatten, die stark genug waren. Und es wäre auch ein Fehler zu sagen, Viren gibt es nicht, nur weil sie mit Lichtmikroskopen nicht zu sehen sind. All das sind Beispiele für Beta-Fehler. Sie alle demonstrieren: je kleiner der Effekt, umso größer der Aufwand, den wir treiben müssen. Wie groß der Effekt ist bestimmt die Realität. Denn die Effektgröße ist eine empirische Größe, so ähnlich wie die Größenordnung eines Pathogens eine empirische Größe ist. Entweder ermitteln wir sie. Das tut man gewöhnlich mit Pilotstudien. Oder man geht von einer vernünftigen klinischen Annahme aus und überlegt sich, wie groß der Effekt in einem konkreten Fall sein muß, damit er den Aufwand wert ist.
Wie groß der Aufwand ist, den wir treiben müssen, um einen Effekt definierter Größe sichtbar machen zu können, bestimmt die Power-Analyse. Sie gibt uns vor, wieviele Patienten wir benötigen, wenn wir in einer Studie einen Effekt einer definierten Größe sichtbar machen wollen und dabei einen alpha-Fehler einer bestimmten Grössenordnung vermeiden wollen – also in der Regel nur in 5 von Hundert Fällen fälschlicherweise einen Unterschied behaupten – und gleichzeitig auch nur mit einer bestimmten Wahrscheinlichkeit einen Effekt übersehen wollen, also einen Beta-Fehler bestimmter Größe, in der Regel 10 bis maximal 20%, nicht unterschreiten wollen. Anders gesagt: mit einer definierten Wahrscheinlichkeit von 80-90% den Effekt, wenn er denn vorhanden ist, auch wirklich zu sehen.

Daraus ergibt sich die ganz allgemeine Regel:

Jeder Effekt, egal wie groß er ist, kann, wenn er tatsächlich und systematisch vorhanden ist, mit einer Stichprobe, die groß genug ist, sichtbar gemacht werden.

Es versteht sich von selbst dass „Größe der Stichprobe“ gleichbedeutung mit „Geld und Aufwand“ ist. Denn die Kosten einer Studie sind direkt proportional zur Anzahl der benötigten Patienten.

Die Größe der benötigten Stichprobe hängt also davon ab,

- wie groß der Effekt ist
- wie sicher ich sein will, dass ich keine falschen Behauptungen über einen vorhandenen Unterschied aufstelle, also von der Größe des alpha-Fehlers, also dem gewählten Signifikanzniveau
- wie sicher ich sein will, dass ich den Effekt, wenn er vorhanden ist, nicht übersehe, also vom akzeptierten Beta-Fehler oder umgekehrt von der statistischen Mächtigkeit des Tests, die ich anziele.

Weil das alles etwas abstrakt bleibt ohne Zahlen, wollen wir das Ganze jetzt an ein paar Beispielen

illustrieren.

Beispiele

1. *Aspirin und Lipidsenker zur Vorbeugung von Herzinfarkt*

Ein klassisches Beispiel ist die sog. „Physicians' Health Study", die in den 80er Jahren durchgeführt wurde [5]. Ärzte wurden eingeladen, an einer Primärpräventionsstudie teilzunehmen, bei der unterschiedliche präventive Maßnahmen verblindet und über längere Zeit getestet wurden. U.a. untersuchte man, ob man mit Aspirin Herzinfarkt vermeiden kann. Die Idee dahinter ist einfach: Aspirin ist ein klassischer Blutverdünner und Entzündungshemmer. Die pathologischen Prozesse die zu Herzinfarkt führen, sind Entzündungsprozesse in der Gefäßwand und vergleichsweise viskoses, also zähflüssiges Blut, und aus der Kombination beider Prozesse ergibt sich eine Minderdurchblutung lebenswichtiger Muskelareale im Herzen. Wenn man sowohl die Entzündungsprozesse hemmen als auch die Viskosität des Blutes verbessern könnte, müsste sich die Herzinfarktwahrscheinlichkeit reduzieren in der Gruppe, die Aspirin vorbeugend nimmt – daher „Primärprävention"; denn die Ärzte waren nicht krank. Man rechnete mit einem kleinen Effekt und so erhielten 11'037 Ärzte Aspirin und 11'034 erhielten Placebo. Die Studie wurde nach etwa 5 Jahren durch das Monitoring-Board, also einer Gruppe, die die Daten während der Studie überwacht, abgebrochen. Denn es hatte sich gezeigt, dass die Aspirin-Gruppe signifikante Vorteile gegenüber der Placebo-Gruppe hatte. Weniger Herzinfarkte waren in der Aspirin-Gruppe aufgetreten, nämlich 47% Risikoreduktion (Das relative Risiko war 0.53, wodurch sich die 47% Reduktion des Risikos berechnet). Das alpha-Niveau von 5% war weit unterschritten worden: der p-Wert lag bei $p < 0.0001$ für Herzinfarkt. Man machte also in weniger als 1 von 10.000 Fällen einen Fehler, wenn man behauptete, Aspirin kann Herzinfarkt verhindern helfen. Ein erstaunt-erschrockener Ausruf ging damals durch die Fachpresse und die Laienblätter. Plötzlich sah es so aus, als müsste alle Welt Aspirin zum Frühstück nehmen. Tatsächlich? Wie groß war der Effekt?
Es waren insgesamt 104 Fälle von Herzinfarkten, davon 5 tödliche, in der Aspirin-Gruppe aufgetreten und 189 Fälle von Herzinfarkt, davon 18 tödliche, in der Kontrollgruppe aufgetreten. Das wird erst dann wirklich gut verständlich, wenn man die absoluten Zahlen sieht: 104 von 11.037 Personen in der Verumgruppe hatten einen Herzinfarkt, also 9 Promille und 189 von 11.034 in der Placebo-Gruppe, also 1.7%. Setzt man 9 Promille zu 1.7% ins Verhältnis (0.9/1.7) erhält man das im Text angegebene relative Risiko von 0.529 oder 0.53. Im Vergleich zur Kontrollgruppe zeigt die Aspirin-Gruppe 47% weniger Herzinfarkte. Da es aber insgesamt nur 1.7% Herzinfarkte gab, ist der eigentliche absolute Effekt sehr klein.
Man kann diese dichotome Effektgröße auch in eine kontinuierliche Effekgröße d transformieren [6] und erhält dann das Äquivalent von $d = 0.05$. Bedenkt man, dass NICE, der englische Regulator, die klinische Relevanzgrenze für viele therapeutische Verfahren, etwa für die Wirksamkeit von Antidepressiva, auf $d = 0.5$ festgesetzt hat, dann ist diese Effektgröße winzig. Ist sie den Aufwand wert? Offenbar nicht. Denn der zwar vorhandene, aber kleine Effekt, wird erkauft durch Nebenwirkungen. So treten etwa unter Aspirin mehr Gehirnblutungen, Schlaganfälle und derlei Dinge auf, was den positiven Effekt des Aspirins wieder aufwiegt. Daher ist Aspirin zur *primären* Prävention von Herzinfarkt wieder aus der Mode gekommen, wird allerdings zur sekundären Prävention, also bei Patienten, die bereits einen Herzinfarkt hatten, durchaus noch eingesetzt.
Dies ist ein klassisches Beispiel dafür, dass ein relativ kleiner Effekt mit einer sehr großen Stichprobe dennoch mit kleinem Beta-Fehler, also mit einer geringen Wahrscheinlichkeit dafür übersehen zu werden, statistisch signifikant „gemacht" werden kann, also wissenschaftlich mit einer relativ geringen Fehlerwahrscheinlich als vorhanden demonstriert zu werden. Die Frage, die sich dann stellt ist: Wie

nützlich im klinischen Sinne ist dieser Effekt? Profitieren ausreichend viele Personen ausreichend stark, so dass man es verantworten kann, in einem gesellschaftlichen und klinischen Sinne, all diejenigen mit zu behandeln, die am Ende gar nicht profitieren werden? Oder nochmals anders: Ist das Verhältnis von Gewinn und Kosten – finanziellen Kosten und Nebenwirkungen – dergestalt, dass es sinnvoll ist, das wissenschaftlich abgesicherte Ergebnis auch wirklich praktisch zu implementieren? Das Aspirin-Beispiel zeigt: es gibt nicht wenige Fälle, bei denen „wisssenschaftlich bewiesen ist, dass" etwas funktioniert oder vorhanden ist, in diesem Fall die Verringerung von Herzinfarkt-Fällen durch die kontinuierliche vorbeugende Einnahme von Aspirin, es aber dennoch Unfug ist, diesen Befund praktisch zu nützen.

Andere Beispiele, für die die Diskussion noch nicht ganz abgeschlossen ist, obwohl man auch darüber trefflich streiten könnte sind die großen Primärpräventionsstudien zur Wirksamkeit von Lipisenkern, gut zusammengefaßt und kritisch diskutiert von Penston [7]. Die großen Studien untersuchten zwischen 4.100 und 19.300 Patienten über mehrere Jahre und untersuchten das Auftreten von Mortalität, oder Herzinfarkt, oder Schlaganfall. Der Prozentsatz der Patienten, die in der Behandlungsgruppe waren und gegenüber der Placebo-Gruppe einen Erfolg hatten, also keines der untersuchten Ereignisse, lag je nach Studie und Ergebnisparameter zwischen 1.4 und 3.8%. Anders ausgedrückt: zwischen 96.2% und 98.8% der Patienten wurden behandelt, ohne dass die Behandlung nötig gewesen wäre bzw. einen sichtbaren Erfolg gehabt hätte, einfach weil die Gesamtzahl der Patienten, bei denen überhaupt ein solches Ereignis wie Schlaganfall oder Herzinfarkt aufgetreten ist, sehr klein ist. Wenn also, wie in der GISSI 3-Studie von mehr als 19.000 Patienten 1.4% profitieren, dann ist das ein sehr kleiner Effekt. Nehmen wir der Einfachheit halber an, von 20.000 Patienten, von denen die Hälfte Lipidsenker erhält und die andere Hälfte Placebo, haben etwa 3% einen Herzinfarkt oder einen Schlaganfall während der Beobachtungszeit, dann wären das 600 Patienten, also im Durschnitt 300 pro Gruppe. Nehmen wir an, eine RR = 0.5, also ein Patient der Behandlungsgruppe hätte eine halb so hohe Wahrscheinlichkeit unter Lipidsenker-Therapie einen Herzinfarkt zu bekommen, dann wären das 400 in der Kontrollgruppe und 200 in der Behandlungsgruppe [8]. Das ist also doppelt so viel und klingt nach viel. Absolut gesehen, ist der Unterschied jedoch sehr klein: denn 200 von 20.000 sind 1% der Stichprobe und 400% sind 2%, zusammen 3% oder 600. Man muß also insgesamt 19.400 Leute mit Lipidsenkern behandeln, um einen solchen Effekt zu erzielen, die in keiner Weise davon profitieren, weil bei ihnen kein Herzinfarkt aufgetreten wäre. So sind diese Zahlen zu lesen. Anders ausgedrückt, hier wird ein winzig kleiner Effekt mit einem riesigen Aufwand von Geld statistisch signifikant gemacht, ohne dass man sich wirklich Rechenschaft darüber abgibt, wie bedeutsam der Effekt ist, ob er im Verhältnis zu Problemen und Nebenwirkungen und den Kosten steht. Lipidsenker sind zwar in der Einzeldosis billig, in der Menge jedoch teuer. Außerdem hemmen sie nicht nur die unerwünschte Fettsynthese im Körper, sondern erniedrigen auch Fettspiegel von essentiellen Fettsäuren und dem wichtigen Coenzym Q10, das wichtig für die Zellfunktion ist, weswegen sie auch sehr häufig sehr schmerzhafte Nebenwirkungen wie Muskelschmerzen verursachen. Daher könnte man in der Tat trefflich streiten, ob der gefundene Effekt tatsächlich klinisch brauchbar ist, zumal man eine Kontrolle der Blutlipide kostengünstig, nebenwirkungsfrei und absolut dauerhaft durch Bewegung und Ernährungsumstellung erreichen kann.

Wir können das an einem aktuellen Beispiel nochmals nachvollziehen, an der JUPITER-Studie, einer der größten bisher durchgeführten Lipidsenker-Primärpräventionsstudien, und zwar an der – verhältnismässig gut ausgegangenen - Auswertung der Frauenkohorte [9]. Die ganze Studie wurde übrigens heftigst debattiert und einige Autoren haben den Autoren der JUPITER-Studie etwas verklausuliert Betrug vorgeworfen [10], aber das wollen wir jetzt nicht weiter verfolgen.

In dieser Studie wurden insgesamt 17.802 Patienten eingeschlossen, davon 6.801 Frauen, um die es hier allein geht. Sie hatten ein unspezifisch erhöhtes C-reaktives Protein. Dies ist ein unspezifischer

Entzündungsmarker und diente in dieser Studie dazu, potenziell gefährdete Personen der Gesamtpopulation aufzufinden, damit die mögliche Trefferrate der Studie höher ist. Die absoluten Auftretensraten für koronare Herzkrankheiten bei den eingeschlossenen Frauen mit erhöhtem C-reaktiven Protein betrugen 0.56 in der Lipidsenker-Gruppe und 1.04 und in der Placebo-Gruppe, standardisiert auf 100-Personen-Jahre (das macht man, weil die Beobachtungszeiten je unterschiedlich sind). D.h. wenn man 100 Menschen ein Jahr beobachtet, dann treten unter dem Lipidsenker ziemlich genau halb so viele Fälle von koronarer Herzkrankheit auf wie unter Placebo.
Auch diese Studie wurde vom Trial-Monitoring Board nach 1.9 Jahren vorzeitig beendet, weil die Behandlungsgruppe statistisch bedeutsame Erfolge gegenüber Placebo erzielt hatte, also die vorher spezifizierte Signifikanzgrenze von 5% erreicht worden war, obwohl man zunächst gedacht hatte, man müsste 5 Jahre beobachten. Der formale Endpunkt der Studie, also das Zielkriterium war ein sog. zusammengesetzter Endpunkt (composite endpoint), in dem alle Fälle von Herzinfarkt, Schlaganfall, Notwendigkeit von Krankenhausaufenthalt oder Operation zur Wiederherstellung von Durchblutung oder Tod eingingen. Das waren insgesamt 39 in der Behandlungsgruppe von 3.426 Frauen und 70 von 3.375 Frauen in der Placebo-Gruppe [11], also etwas mehr als doppelt soviele Frauen in der Kontrollgruppen erlitten eines der voraus definierten Krankheitsereignisse.
Man sieht: die Angabe, ein Ereignis sei in der Behandlungsgruppe nur halb so oft aufgetreten, oder die Behandlungsgruppe sei der Kontrollgruppe ums Doppelte überlegen hat nur dann einen wirklichen Aussagewert, wenn er durch die Gesamtzahl der Ereignisse spezifiziert wird. Allerdings ist die Inzidenz, also das Auftreten des Ereignisses ungefähr nur 1%, und doppelt so hoch in der Kontrollgruppe, also 2%. Während in der öffentlichen Darstellung immer mit *relativen* Erfolgen operiert wird - „doppelt so gut wie in der Kontrolle“, „nur halb so viele Todesfälle, Herzinfarkte, Schlaganfälle mit Lipidsenkern“ - ist diese Aussage nur dann bewertbar, wenn man weiss, wie hoch die *absolute* Zahl des Auftretens war, ein Faktum, das im Übrigen im öffentlich zugänglichen Abstract der JUPITER-Studie geflissentlich verschwiegen wird.
Eine Power-Analyse erübrigt sich hier, da die Studie ja vorzeitig abgebrochen wurde. Was sich nicht erübrigt ist die Überlegung, ob es sinnvoll ist, wegen der 41 von 70 Frauen in der Kontrollgruppe, deren Krankheitsereignis hätte verhindert werden können, insgesamt 6.800 andere zu behandeln, bei denen nie irgendeines der Ereignisse während der Beobachtungszeit aufgetreten wäre. Was passiert, wenn Statine zu einem täglichen Lebensmittel werden, weil alle Welt glaubt, ihre primärpräventologische Wirksamkeit sei wissenschaftlich erwiesen? Welche Nebenwirkungen oder langfristigen Probleme anderer Art handeln wir uns ein? Vielleicht Alzheimer, weil sich die Balance der essentiellen Fettsäuren verschiebt? Vielleicht chronische Müdigkeit, weil die Zellfunktionen durch die Reduktion von Q10 reduziert werden, oder chronische Schmerzsyndrome? Vielleicht ganz andere Probleme, die wir noch nicht im Blick haben? Wir wissen es nicht. Denn solche Fragen werden auch von relativ lang dauernden klinischen Studien nicht untersucht und können daher auch nicht beantwortet werden.
Man sieht: auch winzige Effektstärken können mit entsprechend viel Geld, viel Menschen und langer Beobachtungsdauer signifikant werden und dies mit entsprechender statistischer Mächtigkeit, also einer großen Wahrscheinlichkeit, dass der Effekt sichtbar wird und einer geringen Gefahr, dass er übersehen wird. Nur: Ist er auch nützlich? Wollen wir ihn? Wollen wir dafür bezahlen? Diese Fragen beantworten keine Studien, sondern nur der informierte Diskurs.

2. *Manuelle Therapie bei komplettem Kreuzbandriss*

Die Kreuzbänder stabilisieren das Knie. Häufig führen Sportverletzungen, etwa beim Schifahren, zum Abriß der Kreuzbänder, einer sehr schmerzhaften Verletzung. Spontan heilen solche Verletzungen in vielleicht zwei Jahren komplett aus, wie man aus neueren Studien weiß [12]. Therapeutisch wird

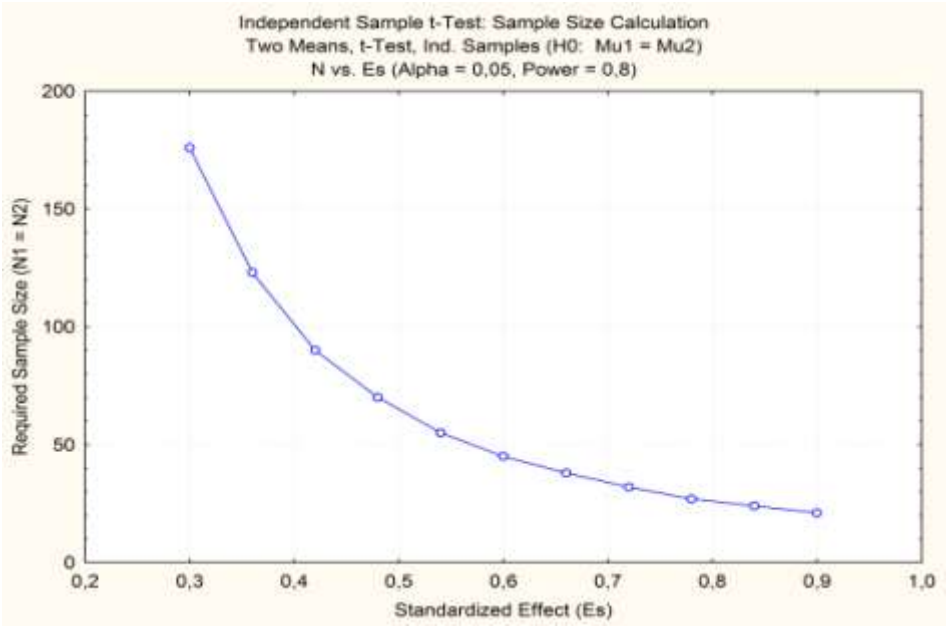
allerdings meistens operiert. Das ist mittlerweile Routine. Allerdings ist es eine teure Routine und eine, bei der es nicht selten zu Komplikationen kommt, etwa Infektionen oder Wundheilungsproblemen, oder vielleicht nach Jahren zu Folgeproblemen wie Arthrosen. Daher haben Spitzensportler schon immer nach Alternativen Ausschau gehalten. In Hallein bei Salzburg gibt es einen bemerkenswerten Manualtherapeuten, Mohammed Khalifa. Jeden Tag widmet er etwa 4 Stunden der Vorbereitung auf seine therapeutische Arbeit, durch physische und geistige Übung. Er behandelt Patienten mit kompletten Bänderrupturen allein mit den Händen. Das ist extrem schmerzhaft, aber sehr wirkungsvoll, so wirkungsvoll, dass sich viele Sportlergrößen, von Spitzenfussballern der Bundesliga bis zu Tennisweltmeistern und Spitzenkletterern bei ihm behandeln ließen. Denn nach der Behandlung sind sie trotz kompletter Ruptur innerhalb von einigen Tagen wieder einsatzfähig.
Wir wollten das untersuchen und haben eine klinische Untersuchung an 30 Patienten mit kompletter Kreuzbandruptur durchgeführt [12, online unter http://dx.doi.org/10.1155/2014/462840]. Die Patienten erhielten randomisiert entweder die Khalifa-Therapie oder konservative, aber sehr gute Physiotherapie. Das Zielkriterium war ein durch den Arzt erfasster Knie-Bewertungswert, der sog. IKDC-Wert; dieser erfasst die objektiv beurteilbare Funktionsfähigkeit des Knies. Zusätzlich verifizierten wir die Therapie-Effekte durch Bildgebung, die zeigen konnte, ob nach 3 Monaten das Band zusammengeheilt war oder nicht. Wir maßen den Erfolg der Therapie am Tag nach der Behandlung und 3 Monate später. Zu diesem Zeitpunkt erhielten auch alle Patienten wieder ein MRI zur Bildgebung. Nach 3 Monaten war bei 7 der 15 Patienten in der Behandlungsgruppe und bei keinem der Patienten der Kontrollgruppe das Kreuzband wieder hergestellt. Das Hauptzielkriterium, der IKDC-Wert, zeigte schon am Tag nach der Behandlung und 3 Monate später sehr starke und signifikante Effekte. War der Effekt nicht nur statistisch signifikant, sondern auch klinisch bedeutsam? War es ein Zufallsbefund oder wie groß war die statistische Mächtigkeit?
Die in der Originalarbeit abgebildete Grafik und Ergebnistabelle geben darüber Auskunft, und wer aufgepasst hat, kann die Analysen auch selber nachvollziehen. Die Effektgröße unmittelbar nach der Behandlung ist mit d = 1.77 ziemlich groß und nach 3 Monaten mit d = 1.19 immer noch sehr groß. Die Behandlungsgruppe liegt also über eine Standardabweichung über der Kontrollgruppe, nach einer einzigen Behandlung. Die statistische Mächtigkeit der Analyse lag mit 0.88 in einem durchaus guten Bereich, auch mit nur 15 Patienten pro Gruppe. Ist das nun ein Effekt, der bedeutsam ist? Wir haben nicht gegen OP verglichen. Schätzungweise würde man gegen OP langfristig die selbe funktionelle Besserung sehen, vermutlich aber nicht kurzfristig, weil durch die OP-Nachbeschwerden zunächst die Funktionstüchtigkeit des Knies mehr eingeschränkt wäre. Die OP kostet mehr Geld und birgt mehr Risiken. Das Risiko der Khalifa-Behandlung ist der einmalige Schmerz und eine noch nicht durch Langzeitstudien ausgelotete Gefahr, dass die Therapie bei manchen erfolglos bleibt. Man muß also auch hier Vor- und Nachteile in die Betrachtung mit einbeziehen und die Effektgröße anhand der vorhandenen Alternativen, ihrer Vor- und Nachteile beurteilen.
Man sieht, auch mit einer vergleichsweise kleinen Studie von insgesamt 30 Teilnehmern lässt sich ein Effekt, wenn er groß genug ist, gut sichtbar machen.

Eigene Übungen:
Es gibt ein hübsches frei verfügbares Programm, G-Power. Damit kann jeder selbst Power-Analysen durchfuhren und das, was ich hier gesagt habe, nachvollziehen http://www.gpower.hhu.de/

Hier sind zwei Grafiken, die anschaulich machen, wie Effektgröße, Stichprobengröße und Power zusammenhängen. Ich habe sie mit meinem STATISTICA-Programm erzeugt, aber auch G-Power bietet die Möglichkeit, solche Grafiken zu erzeugen.
Man sieht sehr leicht: Wenn man eine vernünftige statistische Mächtigkeit anstrebt, 0.8 wird in der Regel als brauchbar und als Minimum angesehen und man hat damit eine Chance, einen Effekt in 80% aller Fälle zu entdecken, dann benötigt man abhängig von der vorhandenen Effektgröße unterschiedlich

viele Patienten. Für den von uns gefundenen Effekt von mehr als einer Standardabweichung war die Power auch für unsere kleine Studie sehr gut. Studien, die häufig gemacht werden, mit etwa 30 Personen pro Gruppe, haben mit anständiger Power nur dann eine Chance, den Effekt zu finden, wenn er größer als d = 0.75 ist, wie man der Grafik entnehmen kann. Das sind etwa Effekte, wie man sie von einer guten Psychotherapie erwarten kann. Zum Vergleich: die adjustierte Effektgröße von Antidepressiva beträgt d = 0.38 [13], etwa gleich groß wie sie von einer sehr konservativ vorgehenden Meta-Analyse für die Behandlung der Angst oder Depression durch Meditation geschätzt wurde [14]. Mittlere Effekte von etwa d = 0.5 oder kleinere erfordern mindestens 70 Leute pro Gruppe und wenn die Effekte klein sind, etwa d = 0.3, muß man 180 pro Gruppe rekrutieren. Umgekehrt kann man schliessen: wenn jemand 5.000 Leute pro Gruppe braucht, dann haben wir es mit winzigen Effekten zu tun.

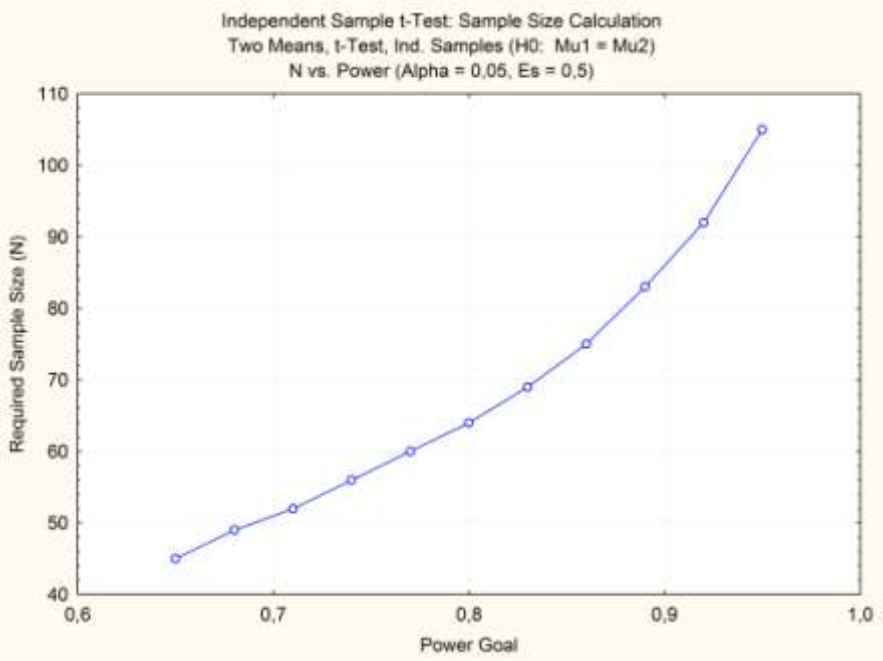
Abb. 1 – Der Zusammenhang von Stichprobengröße und Effektgröße bei einer angenommenen Power



von 0.8

Analog sieht man in Abbildung 2: Wenn man einen hier als fix angenommenen Effekt von einer halben Standardabweichung, also d = 0.5 finden will – das ist der Effekt, den NICE als klinisch relevant bei einer Depressionsbehandlung postuliert -, dann muß man mit steigender Power oder geringerer Bereitschaft, einen Effekt zu übersehen immer mehr Probanden rekrutieren.

Abb. 2 – Der Zusammenhang von Power und Stichprobengröße bei einer angenommenen Effektgröße von d = 0.5



Wir selber sind auch schon mit diesem Problem konfrontiert worden. Wir haben die wohl größte und vielleicht sogar erste aktiv kontrollierte Studie von Achtsamkeitsmeditation zur Behandlung der Fibromyalgie durchgeführt, eine dreiarmige Studie, bei der wir MBSR, eine Form des Gruppentrainings für Achtsamkeit, mit einer aktiven Kontrollbedingung, Stretching und Entspannung, und mit einer Wartegruppe verglichen haben [15]. Wir sind von einer ungefähren Effektstärke von d = 0.55 ausgegangen, die sich aus unserer früheren Meta-Analyse ergeben hat [16]. Dummerweise haben wir die Stärke der aktiven Kontrolle unterschätzt. Obwohl die Achtsamkeitsbedingungen einigen Vorteil bot, das zeigte sich vor allem auch in den Interviews, konnten wir den Effekt nicht statistisch gegen die aktive Kontrolle absichern. Denn die Effektgröße war nur etwa d = 0.3 und aus Abbildung 1 sieht man, dass man etwa 180 Patienten *pro Gruppe* braucht, um einen Effekt dieser Größe zu sichern, unsere Power war also zu gering. Das war etwa die Anzahl der Patienten, die wir für die gesamte Studie hatten. Ist nun deshalb Achtsamkeit bei Fibromyalgie unbrauchbar? Vermutlich nicht. Denn auch eine Effektstärke von nur einem Drittel einer Standardabweichung ist bei dieser Erkrankung ein Erfolg. Denn Fibromyalgie ist schwer bis gar nicht behandelbar und oftmals für die Patienten eine lebenslange Bürde kontinuierlicher Schmerzen. Haben wir gezeigt, dass Achtsamkeit dafür in Frage kommt als Behandlungsstrategie? Nicht im strengen Sinne, denn unsere Studie war nicht signifikant. Das könnte jetzt eine größere Studie versuchen nachzuholen. Dazu müsste jemand wesentlich mehr Geld als die uns zur Verfügung stehenden Mittel einwerben; unwahrscheinlich, wenn man die Datenlage ansieht. Oder aber jemand könnte die Effektstärken akkumulieren über verschiedene Studien hinweg.

Das wäre das Ziel und potenzielle Ergebnis einer Meta-Analyse, und die wird uns in einem der nächsten Blogs beschäftigen.

[1] Fisher, R. A. (1971 (orig. 1935)). *The Design of Experiments*. New York: Hafner.

[2] ein kleines numerisches Beispiel gibt folgende Tabelle:

	behandelt	unbehandelt
Erfolg (a)	40	20
Misserfolg (a')	20	40
Gesamtzahl	60	60

RR = a/ n1 : a'/n2 = 40/60 : 20/60 = 40:20 = 2; bedeutet: zweimal so viele Patienten in der behandelten Gruppe hatten Erfolg. Man beachte: die Gesamtzahl, also den jeweiligen Nenner der Verhältniszahlen kann man ignorieren, wenn die Gruppengröße gleich ist, weil sich die Zahlen dann herauskürzen. Daher kann man einfach 40:20 rechnen.

[3] Welchen Mittelwert man zuerst setzt, ob den der Kontrollgruppe oder den der Behandlungsgruppe ist Frage der Definition und des Geschmacks und hängt auch von der Polung der verwendeten Maße ab. In der Regel formuliert man die Gleichung so, dass ein positiver Wert für d eine Effektgröße im Sinne der Behandlung definiert. Wenn man etwa Depressionswerte erfasst oder Blutdruck, bei denen eine hoher Wert viel Depression bzw. hohen Blutdruck bezeichnet, dann muß man den Wert der Behandlungsgruppe von jenem der Kontrollgruppe subtrahieren. Umgekehrt macht man es etwa bei Lebensqualitätsskalen, die in der Regel so konstruiert sind, dass ein hoher Wert erwünscht ist. Man muß darauf achten, dass man dies bei allen untersuchten Werten konsistent tut.
Die Standardabweichung, die man in den Nenner setzt ist in der Regel die gemittelte, oder „gepoolte" Standardabweichung der beiden Gruppen. Man kann auch, wenn man robust schätzen will, die je größere Standardabweichung verwenden. Denn je größere die Standardabweichung, desto kleiner wird d.
[4] Die Grundlagen der Power-Analyse wurden herausgearbeitet von
Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum. (Original erschienen 1977: New York: Academic Press).
[5] Steering Committee of the Physicians' Health Study Group (1988). Preliminary report: findings from the Aspirin component of the ongoing physican's health study. *The New England Journal of Medicine, 318*, 262-264.
[6] Das ist ein Verfahren, mit dem man Studien, die auf unterschiedlicher Metrik beruhen, approximativ vergleichbar machen kann. Man kann es verwenden, wenn man etwa eine Meta-Analyse durchführt. Entsprechende Handbücher wie Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research*. Newbury Park: Sage enthalten die entsprechenden Angaben. In diesem Falle nimmt man eine Arcsinus-Transformation. Sie transformiert Rate-Ratios in d-Werte, jedenfalls approximativ. Das gleiche erreicht man über die Formel von Hasselblad, V., & Hedges, L. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167-178.
[7] Penston, J. (2003). *Fiction and Fantasy in Medical Research: The Large-Scale Randomised Trial*. London: The London Press.
[8] Wir sehen an Formel [2] für die Rate-Ratio, dass sie im Prinzip unabhängig ist von der Gesamtzahl. Denn wenn die Gesamtzahlen in beiden Gruppen exakt gleich ist, dann kürzen sie sich heraus und übrig bleibt das absolute Verhältnis von Erfolgen in der einen zu Erfolgen in der anderen Gruppe. Nur wenn die Gruppen unterschiedlich groß sind adjustiert die Formel die proportionalen Unterschiede. D.h. aber auch, dass man immer auch das Auftreten des Ereignisses in der Population im Blick haben muss, um den Effekt einschätzen zu können.

[9] Mora, S., Glynn, R. J., Hsia, J., MacFadyen, J. G., Genest, J., & Ridker, P. M. (2010). Statins for the primary prevention of cardiovascular events in women with elevated high-sensitive c-reactive protein or dyslipidemia - Results from the justification for the use of statins in prevention: An intervention trial evaluating rosuvastatin (JUPITER) and meta-analysis of women from prmary prevention trials. *Circulation, 121*, 1069-1077.
[10] de Lorgeril, M., Salen, P., Abramson, J., Dodin, S., Hamazaki, T., Kostucki, W., et al. (2010). Cholesterol lowering, cardiovascular diseases, and the Rusvastatin-JUPITER controversy: A critical reappraisal. *Archives of Internal Medicine, 170*, 1032-1036.
[11] Weil die Gruppen annähernd gleich groß sind, kann man für die Auftretenshäufigkeit die Gruppengrösse vernachlässigen und rechnet HR (Hazard Ratio) = 39/70 = 0.56.
[12] Weitere Literatur, auch zum verwendeten Maß und Spontanheilungsraten in unserer Originalstudie: Ofner, M., Kastner, A., Wallenboeck, E., Pehn, R., Schneider, F., Groell, R., et al. (2014). Manual Khalifa therapy improves functional and morphological outcome of patients with anterior cruciate ligament rupturein the knee: A randomized controlled trial. *Evidence Based Complementary and Alternative Medicine, Art ID 462840*. http://dx.doi.org/10.1155/2014/462840
[13] Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252-260.
[14] Goyal, M., Singh, S., Sibinga, E. M., Gould, N. F., Rowland-Seymour, A., Sharma, R., et al. (2014). Meditation programs for psychological stress and well-being: A systematic review and meta-analysis. *Journal of the American Medical Association - Internal Medicine, doi: 10.1001/jamainternmed.2013.13018*.
[15] Schmidt, S., Grossman, P., Schwarzer, B., Jena, S., Naumann, J., & Walach, H. (2011). Treating fibromyalgia with mindfulness-based stress reduction: Results from a 3-armed randomized controlled trial. *Pain, 152*, 361-369
[16] Grossman, P., Schmidt, S., Niemann, L., & Walach, H. (2004). Mindfulness based stress reduction and health: A meta-analysis. *Journal of Psychosomatic Research, 37*, 35-43.