Folge 19:
Wie kann man Behandlungseffekte aus klinischen Studien bewerten? I: Kontinuierliche Outcomes[1] *

R. D. Herbert

Gut durchgeführte randomisierte kontrollierte Studien (RCTs) können klinische Entscheidungen unterstützen, indem sie unverzerrte Einschätzungen der durchschnittlichen Behandlungseffekte ermöglichen. Diese Folge diskutiert, wie Leser klinischer Studien einfache Schätzungen der Größe von Behandlungseffekten aus den Berichten über Studien heraus ziehen können, wenn die Studienergebnisse kontinuierlich gemessen wurden. Bei Entscheidungen über Behandlungen von individuellen Patienten können diese Schätzungen aufgrund der individuellen Charakteristika eines jeden Patienten modifiziert werden. Modifizierte Einschätzungen der Größe des Behandlungseffekts können benutzt werden, um festzustellen, ob der Behandlungseffekt wahrscheinlich groß genug sein wird, um klinisch gesehen „der Mühe Wert", um lohnenswert zu sein. Dieser Ansatz optimiert klinische Entscheidungsprozesse (clinical decision making), indem er unverzerrte Schätzungen über die Größe von Behandlungseffekten, wie sie sich aus klinischen Studien ergeben mit klinischer Intuition und Präferenzen des Patienten kombiniert.

Einleitung

Randomisierte kontrollierte Studien (RCTs) und systematische Reviews von RCTs haben das Potenzial, unverzerrte und präzise Schätzungen der Wirkungen von Therapien zu erbringen. Aus diesem Grunde können sie, insofern vorhanden, als Schiedsrichter in der Frage fungieren, welche Interventionen als wirksam und welche als nicht wirksam anzusehen sind. Leider sind nicht alle klinischen Studien valide, und die Implikationen valider Studien sind nicht immer offensichtlich, wenn über diese Studien in Artikelform berichtet wird. Daraus folgt, dass, wenn Leser klinischer Studien nicht in die Irre geführt werden sollen, diese eine kritischen Einschätzung (critical appraisal) der Berichte bzw. Artikel über diese Studien vornehmen müssen – und zwar sowohl hinsichtlich der Validität der Studie als auch hinsichtlich dessen, was die Studienergebnisse für die klinische Praxis bedeuten.

Der Prozess der Feststellung der Validität einer Studie ist in vielen Veröffentlichungen dargestellt worden (z. B. *Guyatt* et al 1993, *Sackett* et al 1998)[2]. Er beinhaltet u. a. die Entscheidung, ob die Studie bestimmten methodologischen Kernkriterien gerecht wird, wie z. B. ob eine sauber durchgeführte Randomisierung stattgefunden hat, ob eine angemessene Blindung erfolgt ist und ob ein hinreichender „follow-up", also eine ausreichende Zahl von Messdaten zum zweiten Messzeitpunkt, bzw. eine nur geringe Zahl von Studienabbrechern, erreicht wurde.

√ Der Zweck dieser Folge besteht jedoch darin, einige Punkte, die mit der Interpretation valider Studien einhergehen, darzustellen. Insbesondere wird sie darauf eingehen, wie man entscheiden kann, ob die zu erwartenden Effekte einer Behandlung mit hinreichender Wahrscheinlichkeit groß genug sind, um die Mühen und den Ressourceneinsatz der Intervention lohnenswert erscheinen zu lassen.

Auch diese Fragen wurden bereits an anderer Stelle diskutiert (*Guyatt* et al. 1994, *McAlister* et al 2000, Sackett et al 1998); sie haben jedoch in der physiotherapeutischen Literatur wenig Aufmerksamkeit erhalten.

Warum müssen wir wissen, wie groß Behandlungseffekte sind?

Oft konzentriert sich die Aufmerksamkeit von RCTs auf den „p-Wert"

Die Übersetzung der Originalarbeit (Herbert RD 2000. How to estimate treatment effects from reports of clinical trials. I: Continuous Outcomes. Australian Journal of Physiotherapy; vol. 46; 229-235) erfolgte mit freundlicher Genehmigung des Autors und der Australian Physiotherapy Association, dem Australischen Physiotherapieverband.

* Übersetzung aus dem Englischen: Erwin Scherfer

1 Der Übersetzer empfiehlt Leserinnen und Lesern, denen diese Materie noch eher neu ist, zunächst die Folgen 1 (März 2003) und 9 (Oktober 2003) dieser Serie zu lesen, da in diesen Folgen grundsätzliche Konzepte wie RCT, Mittelwerte und Messniveaus erläutert werden. Diese Konzepte zu kennen, ist notwendig, um diese Folge verstehen zu können.

2 Eine Darstellung der von PEDro verwendeten Validitätskriterien findet sich in Folge 2 dieser Serie; Anm. d. Übers.

der Differenz zwischen den Gruppen. Der „p-Wert“ wird benutzt, um zu entscheiden, ob die Differenz zwischen den Gruppen wahrscheinlich auf einen Behandlungseffekt zurück zu führen ist oder sich auch zufällig ereignet haben könnte. D. h. „p“ ist die Wahrscheinlichkeit dafür, dass die Differenz zwischen den Gruppen auch rein zufällig entstanden ist. Eine geringe Wahrscheinlichkeit (per Konvention von unter 5 %) bedeutet, dass es unwahrscheinlich ist, dass die Differenz sich auch durch das Walten des Zufalls eingestellt haben könnte, und stellt damit Evidenz für einen Behandlungseffekt dar. Höhere Wahrscheinlichkeiten (per Konvention von 5 % und größer) zeigen an, dass die Differenz auch zufällig hätte auftreten können[3]. Hohe p-Werte werden richtigerweise als Mangel an Evidenz für einen Behandlungseffekt interpretiert. Eine Folge dieser komplizierten Denkweise ist, dass Leser von der wichtigsten Information, die eine klinische Studie bereit halten kann, abgelenkt werden, nämlich von der Information über die Größe des Behandlungseffekts. Wenn klinische Studien die klinische Praxis beeinflussen sollen, müssen sie mehr tun als nur zu entscheiden, ob eine Behandlung eine Wirkung hat oder nicht. Sie müssen zusätzlich ermitteln, wie groß der Behandlungseffekt ist. Gute klinische Studien bieten unverzerrte Schätzungen der Größe des Behandlungseffekts. Solche Schätzungen können genutzt werden, um zu entscheiden, ob von einer Behandlung ein ausreichend großer Effekt erwartet werden kann, d.h. ein Effekt, der in klinischer Hinsicht als „lohnenswert“ betrachtet werden kann.

Was aber ist ein klinisch lohnenswerter Effekt?

Das hängt ab von den Kosten und den Risiken der Behandlung. Kosten beinhalten einerseits monetäre Kosten (für den Patienten, den Gesundheitsdienstleister oder den Kostenträger), aber sie beinhalten genauso die Unannehmlichkeiten, die Beschwerden und Nebenwirkungen von Interventionen. Um klinisch „lohnenswert“ zu sein, müssen die positiven Effekte größer sein als die Kosten; die Behandlung muss mehr nutzen als schaden. Klinische Studien geben zwar oft Informationen über die Größe von Behandlungseffekten, aber sie informieren selten über die soeben definierten Kosten der Behandlungen. Die Beantwortung der Frage, ob eine Behandlung einen lohnenswerten klinischen Effekt hat, erfordert daher normalerweise eine Abwägung von objektiven Informationen über positive Behandlungseffekte (ermittelt über klinische Studien) mit subjektiven Einschätzungen hinsichtlich der Kosten und Risiken einer Behandlung (die nur von Patient und Therapeut erbracht werden können).

Was können uns Studien über die Wirkungen von Behandlungen sagen?

Die Wirkungen aller Therapien sind variable Größen. Viele haben einen günstigen Effekt auf manche Patienten, aber haben keinen oder sogar schädigenden Einfluss auf andere. Somit können wir streng genommen nicht von *dem* Effekt einer Behandlung sprechen. Aber welche sinnvollen Informationen kann uns dann eine klinische Studie bringen, wenn sie uns nicht sagen kann, ob alle unsere Patienten (oder jeder einzelne Patient) positiv auf die in der Studie untersuchte Behandlung reagieren?

√ Klinische Studien können eine Schätzung der *durchschnittlichen Wirkung* einer Behandlungsmethode erbringen.

√ Nun bedeutet glücklicherweise, über die durchschnittliche Wirkung einer Behandlungsmethode Bescheid zu wissen, auch, über den *wahrscheinlichsten Effekt* einer Behandlung Bescheid zu wissen.

√ Tatsächlich sind die beiden Größen in den meisten Fällen identisch.

Mithin: Während uns klinische Studien nicht sagen können, wie die Wirkung einer Behandlung auf einen individuellen Patienten sein wird, so können sie uns doch sagen, was der wahrscheinlichste Effekt ist. Die gleiche Einschränkung betrifft alle Informationsquellen über Behandlungseffekte; die Einschränkung gilt nicht nur für klinische Studien.

Klinische Studien liefern eine bestmögliche erste Vermutung

Eine sinnvolle Art und Weise, Schätzungen über durchschnittliche Behandlungseffekte, wie sie von klinischen Studien geliefert werden, zu nutzen, ist, sie als eine bestmögliche erste Vermutung oder Erwartung hinsichtlich der wahrscheinlichen Größe

3 Eine Auftrittswahrscheinlichkeit von 5 % oder mehr für die beobachtete Differenz zwischen den Gruppen wird dahingehend interpretiert, dass der Zufall nicht mit hinreichender Sicherheit für die „Erzeugung“ des Unterschieds ausgeschlossen werden kann. Die Differenz zwischen den Gruppen gilt statistisch als „nicht signifikant“. Siehe hierzu auch Folge 2 dieser Serie in der Zeitschrift für Physiotherapeuten, Heft 3, 2003; Anm. d. Übers.

des Behandlungseffekts und damit für die Definition von Behandlungszielen zu verwenden. Diese Vermutung kann dann nach oben oder unten verändert werden, je nach den Eigenschaften des individuellen Patienten, bei dem die Therapie angewendet werden soll. So zeigte eine klinische Studie von *Dean* und *Shepherd* (1998), dass ein zweiwöchiges spezifisches motorisches Training nach Schlaganfall die maximale Greifreichweite im Sitzen um ca. 8 cm verlängert. Die Probanden hatten einen Schlaganfall erlitten, der mehr als ein Jahr zurücklag und litten weder an Demenz noch an Wahrnehmungsaphasie. Wir könnten größere Effekte als die von *Dean* und *Shepherd* festgestellten erwarten, wenn das Training früher nach dem Schlaganfall durchgeführt würde, und kleinere Effekte, wenn die Patienten an Demenz oder Aphasie leiden würden. Dieser Ansatz kombiniert die Objektivität klinischer Studien mit ihren unverzerrten Schätzungen der durchschnittlichen Wirksamkeit von Behandlungsmethoden mit dem Reichtum klinischer Analysefähigkeit (die helfen kann, zwischen wahrscheinlich positiv auf die Therapie ansprechenden und wahrscheinlich nicht positiv ansprechenden Patienten zu unterscheiden.)

√ Natürlich muss man sehr vorsichtig vorgehen, wenn man klinische Intuition gebraucht, um die Einschätzung von Behandlungseffekten, die sich aus klinischen Studien ergeben haben, zu modifizieren.

√ Ein konservativer Ansatz wäre, sicher zu stellen, dass die Größe des Behandlungseffekts ebenso oft nach oben wie nach unten korrigiert wird.

√ Aber es kann auch sinnvoll sein, von diesem Ansatz abzuweichen, wenn sich die Patienten, die an der Studie teilnahmen, deutlich von den behandelten Patienten, also von der klinischen Population, unterscheiden.

√ Mit besonderer Vorsicht ist vorzugehen, wenn eine klinische Studie Evidenz für die Nicht-Wirksamkeit einer Therapie erbringt.

Lohnt sich die Behandlung in klinischer Sicht?

Um diese Frage beantworten zu können, müssen die Behandlungseffekte gegenüber den „Kosten" der Behandlung abgewogen werden. Der erste Schritt im Prozess der Entscheidung, ob von einer Behandlung ein klinisch gesehen lohnenswerter Effekt zu erwarten ist, besteht darin, zunächst den kleinsten klinisch lohnenswerten Effekt zu benennen. Hierbei handelt es sich um eine subjektive Entscheidung, die auch die Wahrnehmung des Patienten bezüglich des Nutzens und der Kosten der Behandlung berücksichtigt. Die meisten Therapeuten ziehen den kleinsten klinisch lohnenswerten Behandlungseffekt in Betracht, wenn sie über die Anwendung oder Nicht-Anwendung einer bestimmten Behandlung entscheiden. Manchmal wird die Festlegung des kleinsten klinisch lohnenswerten Behandlungseffekts explizit mit dem Patienten ausgehandelt.

Beispiel

Dieser Prozess soll am Beispiel der Anwendung einer pneumatischen Kompressionspumpe, eingesetzt mit dem Ziel der klinisch lohnenswerten Reduktion von Lymphödemen nach Mastektomie illustriert werden. Wir könnten beginnen mit einer Quantifizierung der *kleinsten* lohnenswerten Verringerung eines Lymphödems, welche die mit der Behandlung einhergehenden „Kosten" der Behandlung aufwiegen kann. Die meisten Therapeuten, und vielleicht sogar die meisten Patientinnen, würden darin übereinstimmen, dass eine kurze tägliche Anwendung der Kompressionstherapie klinisch lohnenswert wäre, wenn dies eine anhaltende Verringerung des Lymphödems um 75 % bewirken würde. Die meisten würden auch zustimmen, dass eine 15 %ige Verringerung nicht klinisch lohnenswert wäre. Irgendwo zwischen diesen beiden Werten liegt der kleinste klinisch lohnenswerte Behandlungseffekt. Diesen Wert legt man am besten fest, indem man ihn mit derjenigen Patientin, die die Therapie bekommen soll, bespricht. Nehmen wir für dieses Beispiel an, dass eine bestimmte Patientin (oder typische Patientinnen) eine Ödemreduktion von ungefähr 40 % als kleinsten klinischen Effekt betrachten, der die Behandlung lohnenswert macht.

Ist dieser „kleinste" Effekt zu erreichen?

Erreicht Kompressionstherapie eine Reduktion von Lymphödemen in dieser Größenordnung? Die vielleicht beste Antwort auf diese Frage lässt sich einer randomisierten Studie von *Dini* et al. (1998) entnehmen, in der eine zweiwöchige (10 Tage) intermittierende pneumatische Kompression mit Beratung allein verglichen wurde. Wir werden die Ergebnisse dieser Studie nutzen, um zu schätzen, welchen Effekt wir von der Kompressionstherapie erwarten können.

Nehmen wir die Studie von *Dini* et al.

Die beste Schätzung des Behandlungseffekts ist einfach die Differenz

der Mittelwerte (oder in manchen Studien: der Mediane) von Behandlungs- und Kontrollgruppe[4]. In der Studie von *Dini* et al. (1998) wurde der Ödemumfang folgendermaßen gemessen: An sieben Stellen wurde der Armumfang gemessen, die Messwerte wurden summiert und schließlich wurde die Differenz zwischen der Summe der Messwerte des betroffenen Arms und der des nicht betroffenen Arms als Maß des Ödems genommen (wobei positive Zahlen angeben, dass der betroffene Arm einen größeren Umfang hatte als der nicht betroffene Arm). Nach der zweiwöchigen Experimentalperiode betrug der Ödemwert 14,1 cm (SD 5,6 cm)[5] in der Kontrollgruppe und 14,2 cm in der Behandlungsgruppe. Somit ist die beste Schätzung für den Behandlungseffekt, dass er Ödeme um 0,1 cm vergrößert (da 14,1 cm – 14,2 cm = -0,1 cm). Da der Ödemwert vor der Behandlungsperiode im Durchschnitt 15,5 cm betrug, entspricht die Größe des Behandlungseffekts einer Ödemzunahme von weniger als einem Prozent (100 x 0,1/15,5). Dieser Behandlungseffekt ist deutlich kleiner als der kleinste klinisch lohnenswerte Effekt (den wir zuvor auf ca. 40 % Verringerung festgelegt hatten). Tatsächlich geht der Behandlungseffekt sogar in die falsche Richtung, da die behandelte Gruppe – wenn auch nur in geringem Maße – höhere Ödemmesswerte aufwies als die Kontrollgruppe. Wir können erwarten, dass der Behandlungseffekt von Kompressionstherapie in der von *Dini* et al. beschriebenen Weise in dieser Population klein sein wird. Unsere beste Vermutung ist, dass der Therapieeffekt im Durchschnitt (also am wahrscheinlichsten) kleiner sein wird als der als klinisch lohnenswert anzusehende Effekt.

In dem soeben verwendeten Beispiel wurden die Ergebnisse, die *Outcomes*, in Form der Ödemumfangwerte zum Ende der Experimentalperiode gemessen. Andere Studien hingegen geben die Veränderung der Outcomevariable während des Behandlungszeitraums an. Auch in solchen Studien ist das Maß des Behandlungseffekts die Differenz zwischen den Mittelwerten – dann der Differenz zwischen den durchschnittlichen Veränderungen (der Mittelwerte der Veränderungen) zwischen Behandlungs- und Kontrollgruppe.

Quantifizierung von Unsicherheit

Auch wenn klinische Studien gut geplant und durchgeführt wurden, sind ihre Ergebnisse mit Unsicherheiten behaftet. Und zwar deswegen, weil die in der Studie beobachtete Differenz zwischen den Mittelwerten der Gruppen selbst ja nur eine Schätzung des wahren Behandlungseffekts ist, die aus der Stichprobe von 80 Probanden, die an der Studie von *Dini* et al. teilnahmen, geschlossen wurde. Wie in jeder Stichprobe nähern sich die Stichprobenwerte dem wahren Wert des Behandlungseffekts an, aber sie gleichen nicht exakt dem durchschnittlichen Ergebnis der Behandlung für die Population, die durch die Stichprobe repräsentiert werden soll.[6] Die in dieser Studie gemessene Größe des Behandlungseffekts nähert sich also der wahren Größe des Behandlungseffekts an, ist aber nicht mit ihr gleich zu setzen. Eine rationale Interpretation einer klinischen Studie erfordert Überlegungen darüber, wie gut die Annäherung der Schätzung an den wahren Wert ist. D. h.

√ um die Ergebnisse einer Studie angemessen zu interpretieren, muss man wissen, mit wie viel Unsicherheit die Ergebnisse verbunden sind.

Das 95 %-Konfidenzintervall

Der Grad der Unsicherheit, der mit der Größe eines Behandlungseffekts einhergeht, kann mit einem Konfidenzintervall beschrieben werden (*Gardner* und *Altman* 1989; *Sim* and *Read* 1999). Meistens wird hierfür das „95 %-Konfidenzintervall“ benutzt. Grob gesagt, gibt das 95 %-Konfidenzintervall den Wertebereich von Behandlungseffekten an, innerhalb dessen wir zu 95 % sicher sein können, dass er den wahren durchschnittlichen Behandlungseffekt enthält (zu beachten ist, dass das Konfidenzintervall den Grad der Unsicherheit hinsichtlich des Behandlungseffekts bezüglich der Population quantifiziert, nicht aber den Grad der Unsicherheit hinsichtlich des Behandlungseffekts bezogen auf Individuen). Das 95 %-Intervall für die Differenz der Mittelwerte in der Studie von Dini et al. reicht von – 2.9 cm bis 2.7 cm (die Methodik zur Errechnung der Konfidenzintervalle wird weiter unten dargestellt), oder – insofern die Ödeme in Prozent des Ödemumfänge zu Beginn der Studie ausgedrückt werden, von –19 % bis 17 %. Dies zeigt an, dass wir sicher sein kön-

4 Eine detaillierte Darstellung der Logik randomisierter kontrollierter Studien findet sich in Folge 1 dieser Serie (Zeitschrift für Physiotherapeuten; Heft 2, 2003; Mittelwerte und Standardabweichungen sind in den Folgen 9 und 10 (Zeitschrift für Physiotherapeuten; Hefte 10 und 11, 2003) erläutert worden; Anm. d. Übers.

5 SD = Standard Deviation = Standardabweichung.

6 Zur Logik von Population (Grundgesamtheit) und Stichprobe siehe Folgen 13 und 14 dieser Serie (Zeitschrift für Physiotherapeuten, Hefte 2 und 3, 2004; Anm. d. Übers.

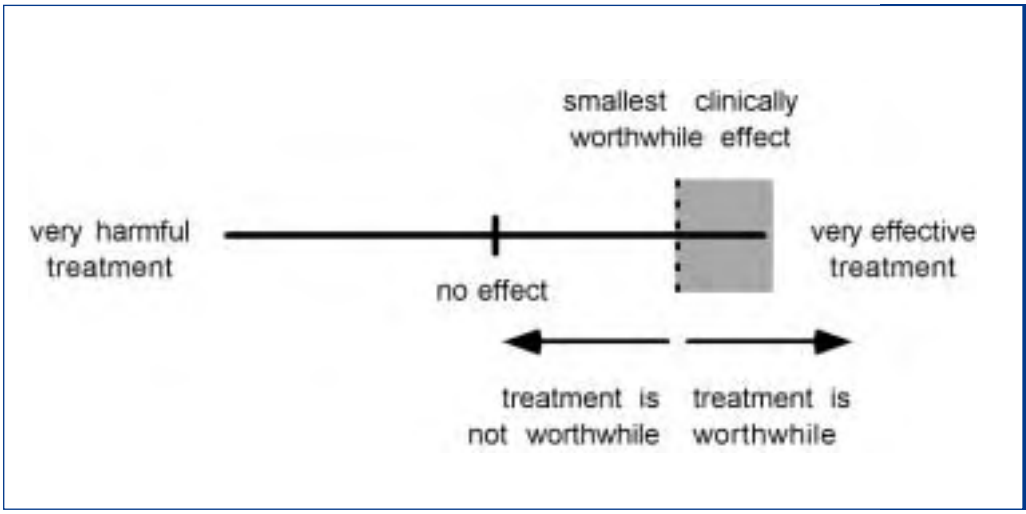

Abb. 1:
Baumdiagramm der Effektgröße. Das Baumdiagramm besteht aus einer horizontalen Linie, auf der die Behandlungseffekte darstellbar sind. An den beiden Enden sind jeweils sehr schädliche und sehr effektive Behandlungen. Der kleinste klinisch lohnenswerte Effekt ist als unterbrochene vertikale Linie eingezeichnet. Der Bereich rechts von der Linie stellt klinisch lohnenswerte Effekte dar

nen, dass der wahre Effekt der Kompressionstherapie irgendwo zwischen einer Ödemumfangszunahme von 19 % und einer Verringerung von 17 % liegt. Alle Werte, die das 95 %-Konfidenzintervall umfasst, sind kleiner als der kleinste klinisch lohnenswerte Effekt. Somit können wir folgern, dass nicht nur die beste Schätzung der Größe des Behandlungseffekts kleiner als der kleinste klinisch lohnenswerte Effekt ist (- 1% < 40 %), sondern auch, dass kein Wert des Behandlungseffekts, der plausibel mit den Ergebnissen der Studie vereinbar ist (indem er innerhalb des Konfidenzintervalls liegt), den kleinsten klinisch lohnenswerten Effekt übersteigt.

√ Diese Daten legen sehr nahe, dass Kompressionstherapie, zumindest so, wie sie von *Dini* et al. angewandt wurde, keine klinisch lohnenden Reduktionen lymphatischer Ödeme erreicht.

Konfidenzintervall als Baumdiagramm

Einige Leser werden es leichter finden, Konfidenzintervalle zu interpretieren, wenn sie sie mit Hilfe eines „Baumdiagramms“, wie in Abbildung 1 gezeigt, grafisch darstellen (Abb. 1). Ein Baumdiagramm besteht aus einer Linie, auf der die sich unterscheidenden Behandlungseffekte angeordnet sind. Die Mitte der Linie stellt „keine Wirkung“ dar (die Differenz zwischen den Mittelwerten der Gruppen beträgt 0). Das rechte Ende der Linie repräsentiert einen sehr guten Behandlungseffekt (die Differenz zwischen dem Mittelwert der Behandlungsgruppe und dem Mittelwert der Kontrollgruppe ist eine große positive Zahl); und das linke Ende der Linie steht für sehr schädliche Behandlungen (Mittelwert der Behandlungsgruppe minus Mittelwert der Kontrollgruppe ist eine große negative Zahl). Für jede Studie können wir drei Variablen in das Diagramm einzeichnen (Abb. 2):

1. Den kleinsten klinisch lohnenswerten Effekt (in unserem Beispiel beträgt dieser 40 %),
2. die beste Schätzung des Behandlungseffekts aus der Studie (die Differenz zwischen den Mittelwerten der Gruppen aus *Dinis* et al. randomisierter kontrollierter Studie, bzw. –1 %) und
3. das 95 %-Konfidenzintervall zu dieser Schätzung (-19 % bis 17 %).

Der Bereich rechts vom kleinsten klinisch lohnenswerten Effekt ist die Zone klinisch lohnenswerter Behandlungseffekte. Die Grafik für die Studie von *Dini* et al. (Abb. 2 B) zeigt deutlich, dass eine klinisch lohnenswerte Behandlungswirkung nicht existiert, weil weder die beste Schätzung des Behandlungseffekts, noch irgendein Punkt des 95 %-Konfidenzintervalls in der Region klinisch lohnenswerter Behandlungseffekt liegt.

Leben mit Unsicherheit

In dem soeben dargestellten Beispiel war der Behandlungseffekt eindeutig nicht groß genug, um als klinisch lohnenswert gelten zu können. Manchmal hingegen wird der Behandlungseffekt sich als eindeutig klinisch lohnenswert erweisen. Sehr oft werden sich die Ergebnisse allerdings weniger klar darstellen. Unklarheit entsteht, wenn das Konfidenzintervall den kleinsten klinisch lohnenswerten Behandlungseffekt umfasst, denn dann ist es gleichermaßen plausibel anzunehmen, dass die Behandlung einen klinisch lohnenswerten Effekt hat, wie auch, dass sie ihn nicht hat (ein Teilbereich der Werte, die das Konfidenzintervall umspannt, liegt über

dem kleinsten klinisch lohnenswerten Effekt, und ein Teil darunter, beide Resultate können also zutreffen). So zeigten z. B. *Sand* et al. (1995), dass eine 15-wöchige Behandlung mit Elektrostimulation des Beckenbodens im Vergleich mit einer Placebo-Behandlung bei Frauen mit idiopathischer Stressinkontinenz eine deutliche Verringerung des unwillkürlichen Urinabgangs (durchschnittlich 32 ml oder 70 % Reduktion) erreicht. In Abbildung 2 B werden diese Ergebnisse in einem Baumdiagramm dargestellt. Die Differenz der Mittelwerte legt einen großen und klinisch lohnenswerten Behandlungseffekt nahe, aber das 95 %-Konfidenzintervall reichte von einer 7 %igen bis zu einer 100 %igen Verringerung. Folglich besteht ein hohes Maß an Unsicherheit darüber, wie groß der Behandlungseffekt wirklich ist, und weil das „untere Ende" des Konfidenzintervalls minimale Verringerungen des unwillkürlichen Urinabgangs einschließt, kann – zumindest auf Basis dieser Studie – nicht sicher davon ausgegangen werden, dass sich die Behandlung lohnt.

Wie erklärt sich das?

Eine solche Situation, in der das Konfidenzintervall den kleinsten klinisch lohnenswerten Effekt umfasst, ist meistens auf einen oder beide der folgenden Gründe zurück zu führen:

Erstens werden viele Studien hinsichtlich ihrer Stichprobengröße so geplant, dass sie zwar in der Lage sind, statistisch signifikante Unterschiede festzustellen, die Konfidenzintervalle aber den kleinsten klinisch lohnenswerten Effekt beinhalten. Die Stichprobenumfänge sind zu klein.

Zweitens haben viele Behandlungen bescheidene Wirkungen (die wahren Effekte sind dicht beim kleinsten klinisch lohnenswerten Effekt), sodass die Konfidenzintervalle sehr klein sein müssen, um diesen nicht einzuschließen. In Folge dessen schaffen nur wenige Studien eindeutige Evidenz für oder gegen einen Behandlungseffekt.

Was also tun?

Es gibt zwei Möglichkeiten, auf die Unsicherheit, die oft durch einzelne Studien bewirkt wird, zu reagieren.

1. Können wir die Unsicherheit akzeptieren und auf der Basis der besten erhältlichen Evidenz mit unserer Praxis fortfahren. Bei diesem Ansatz werden klinische (praktische) Entscheidungen auf Basis der Mittelwertdifferenz zwischen den Gruppen getroffen. Wenn die Differenz den kleinsten klinisch lohnenswerten Behandlungseffekt übersteigt, dann wird angenommen, dass sich die Behandlung lohnt. Ist die Mittelwertsdifferenz kleiner als der kleinste klinisch lohnenswerte Effekt, dann wird der Behandlung mangelnde Wirksamkeit unterstellt. Bei diesem Ansatz bleibt die Rolle des Konfidenzintervalls beschränkt auf die eines Indikators für den Grad des Zweifels, der bestehen bleibt, aber sie berührt die klinische Entscheidung für oder gegen die Therapie anderweitig nicht.

2. Alternativ hierzu besteht die Möglichkeit, Sicherheit zu suchen, indem man festzustellen versucht, ob die Ergebnisse einzelner Studien von anderen, ähnlich angelegten Studien bestätigt werden. Dies ist einer der Hauptgründe, warum systematische Reviews von randomisierten kontrollierten Studien so „populär" geworden sind (*Chalmers* und *Altman* 1995)[7].

Systematische Reviews können ein Ausweg sein

In systematischen Reviews können die Ergebnisse einzelner Studien statistisch in einer Metaanalyse zusammengefasst ausgewertet werden, wodurch letztendlich *ein* Ergebnis aus vielen Studien gewonnen wird. Dieses eine, aggregierte Ergebnis ist dann aus einem relativ großen Sample gewonnen worden, wodurch in der Regel eine präzisere Schätzung der Größe des Behandlungseffekts ermöglicht wird (mit einem relativ kleinen Konfidenzintervall); dadurch wird es wahrscheinlicher, dass eindeutige Informationen über die Größe des Behandlungseffekts gewonnen werden können (engere Konfidenzintervalle gehen mit einer geringeren Wahrscheinlichkeit einher, dass der kleinste klinisch lohnenswerte Effekt durch sie umfasst wird). Ein Beispiel hierfür ist der systematische Review mit Metaanalyse von *Zhang* et al. (1996), der zeigte, dass der Geburtsvorgang bei Erstgebärenden, wenn sie professionelle Unterstützung während der Wehen erfahren, kürzer ist als bei Frauen ohne professionelle Hilfe (aggregierte Mittelwertsdifferenz 2,8 Stunden; 95 %-Konfidenzintervall 2,2 bis 3,4 Std.). Dieser Behandlungseffekt, der in Abbildung 2 C in einem Baumdiagramm dargestellt ist, ist groß, und sogar die pessimistischste Interpretation der Daten (professionelle Hilfe verkürzt den Geburtsvorgang um über zwei Stunden) legt einen klinisch lohnenswerten Effekt nahe.

7 Eine nähere Darstellung zu systematischen Reviews und Metaanalysen findet sich in Folge 17 dieser Serie (Zeitschrift für Physiotherapeuten, Heft 7, 2004); Anm. d. Übers

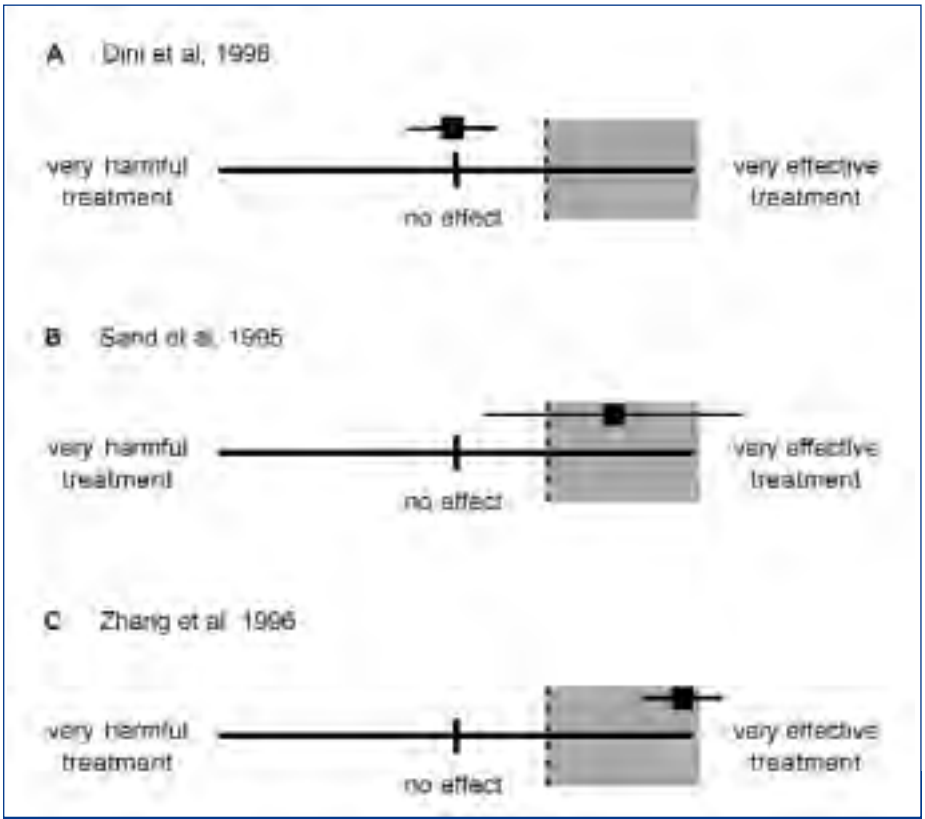

Abb. 2:
(A) Daten von Dini et al. (1998) zur Reduktion von Ödemen. Der kleinste klinisch lohnenswerte Effekt wurde auf 40 % gesetzt. Die beste Schätzung der Größe des Behandlungseffekts (-1 %) ist als kleines Quadrat abgebildet, und das 95 %-Konfidenzintervall zu dieser Schätzung ist als kurze horizontale Linie, aus dem Quadrat heraus, abgebildet. Die Wirkung ist deutlich kleiner als der kleinste klinisch lohnenswerte Effekt.
(B) Daten von Sand et al. (1995) zur Reduktion von Urininkontinenz. Der kleinste klinisch lohnenswerte Effekt wurde auf 40 % gesetzt. Die beste Schätzung der Größe des Behandlungseffektes (70 %) und das dazugehörige 95 %-Konfidenzintervall (7-100 %) sind eingezeichnet. Die beste Schätzung des Behandlungseffekts ist also, dass er klinisch lohnenswert ist, aber gleichzeitig ist diese Schätzung einem hohen Grad von Unsicherheit unterworfen.
(C) Daten von Zhang et al. (1996) zur Verkürzung des Geburtsvorgangs. Der kleinste klinisch lohnenswerte Effekt wurde als eine Stunde bestimmt. In das Diagramm eingezeichnet sind die beste Schätzung des Behandlungseffekts (2,8 Stunden) und das 95 %-Konfidenzintervall (2.2-3.4 Stunden). Diese Behandlung hat einen deutlichen klinisch lohnenswerten Effekt.

Berechnung von Konfidenzintervallen für Mittelwertdifferenzen

Insofern in Berichten über Studien Konfidenzintervalle zu Mittelwertdifferenzen nicht explizit angegeben werden, ist es in der Regel ein Leichtes, diese aus den in der Studie berichteten Daten zu berechnen. Das Konfidenzintervall für die Differenz der Mittelwerte zweier Gruppen kann berechnet werden, wenn folgende Informationen vorliegen: Beide Mittelwerte (bzw. damit die *Differenz* zwischen beiden), ihre Standardabweichungen und die Gruppengrößen (Stichprobenumfänge). Ein angenähertes 95 %-Konfidenzintervall erhält man, indem man zunächst den Durchschnitt der beiden Standardabweichungen (SDs) und den Durchschnitt der Gruppengrößen (n) berechnet. Hieraus lässt sich das angenäherte 95%-Konfidenzintervall (KI) wie folgt berechnen:

$$95\ \%\ KI = Differenz \pm 3 \times SD/\sqrt{n}$$ [8]

In Worten gesagt,
√ umfasst also das Konfidenzintervall eine Spanne von drei Standardabweichungen unterhalb der Mittelwertdifferenz bis drei Standardabweichungen oberhalb der Mittelwertdifferenz

Diese Formel ist eine Annäherung an komplexere Gleichungen, die für die Erstanalyse von Daten in Studien verwendet werden sollten, aber es handelt sich um eine hinreichend genaue Annäherung, wenn es darum geht, durch das Lesen von Studien klinische Entscheidungen zu unterstützen. Sie hat zudem den Vorteil, dass sie einfach genug ist, um routinemäßig berechnet zu werden, wann immer eine Studie kein Konfidenzintervall für die Mittelwertdifferenz zwischen den Gruppen berichtet.

In der Studie von *Dini* et al., an der 80 Probanden teilnahmen (woraus sich eine durchschnittliche Gruppengröße von n = 40 ergibt), berichteten die Autorinnen mittlere Ödemumfänge sowohl für die Behandlungs- als auch für die Kontrollgruppe (14,2 cm bzw. 14,1 cm), und die SDs zu diesen Mittelwerten (6,0 bzw. 5,6 cm; die gemittelte SD beträgt somit 5,8 cm), aber nicht das 95 %-Konfidenzintervall für die Mittelwertsdifferenz. Dieses kann aus den gegebenen Daten wie folgt berechnet werden.

8 Eine Ableitung dieser Annäherung findet sich im Anhang.

95% KI ≈ (14,1 – 14,2) ± 3 x 5,8/√40
≈ -0,1 ± 2,8
≈ -2,9 bis + 2,7cm

Oft geben Artikel über Studien den Standardfehler (standard error, SE)[9] anstelle der Standardabweichung an. In diesem Fall ist die Annäherung sogar noch einfacher:

95% KI = Differenz ± 3 x SE

Viele Studien arbeiten mit mehr als zwei Gruppen, z. B. indem sie mehr als eine Behandlungsgruppe oder mehr als eine Kontrollgruppe aufweisen. Leser müssen dann entscheiden, welche(r) Zwischengruppenvergleich(e) am interessantesten sind. Dann können die Konfidenzintervalle für die Mittelwertsunterschiede auf die gleiche Art und Weise wie soeben dargestellt berechnet werden. Die meisten Studien berichten auch über mehr als ein, und manchmal über viele Outcomes. Es ist eine ermüdende Angelegenheit, Konfidenzintervalle für alle Outcomes zu berechnen, und am besten entscheidet man sich für die Outcomes, die einen am meisten interessieren, und berechnet die 95 %-Konfidenzintervalle nur für diese.

Auf den Spuren von Sherlock Holmes

Manchmal ist ein gewissen Maß detektivischer Kleinarbeit erforderlich, um die SDs oder SEs der Mittelwertdifferenzen zu entdecken. Wenn sie nicht explizit angegeben sind, kön-

9 Der Standardfehler ist ein Streuungsmaß für Mittelwerte und ergibt sich aus der Division der Standardabweichung durch die Wurzel aus dem Stichprobenumfang: SD/√n; Anm. d. Übers.

Anhang:

Angenähertes 95 %-Intervall für die Differenz der Mittelwerte zwischen zwei Gruppen

Die eigentliche Formel für das Konfidenzintervall für die Mittelwertdifferenz zwischen zwei Gruppen ist:

$$KI = Differenz \pm t_{(1\ \ a/2)} \sqrt{\frac{(n_t\ \ 1)SD_t^2 + (n_c\ \ 1)SD_c^2}{n_t + n_c\ \ 2}} \sqrt{\frac{1}{n_t} + \frac{1}{n_c}}$$

Dabei ist *„Differenz“* die Differenz zwischen den arithmetischen Mitteln aus beiden Gruppen, $t_{(1-a/2)}$ der entsprechende Wert einer t-Verteilung, *n* die Probandenzahl und *SD* die Standardabweichung in einer Gruppe, wobei die kleinen Indizes *t* und *c* „der Behandlungsgruppe“ (*t* für treatment) bzw. „der Kontrollgruppe“ (*c* für controlls) bedeuten (*Gardner* und *Altman* 1989). In randomisierten kontrollierten Studien sind die Gruppengrößen in der Regel annähernd gleich groß ($n_t \approx n_c$) und es wird in der Regel angenommen, dass die Varianzen gleich sind (so dass $SD_t = SD_c$). Wenn nun $n_t = n_c = n$, und $SD_t = SD_c = SD$, vereinfacht sich der Ausdruck zu: 95 % KI = *Differenz* $\pm\ t_{(1-a/2)}$ x $\sqrt{2}$ x *SD* / $\sqrt{n}$
Weil $t_{(1-a/2)}$ für das 95 %-KI ≈ 2, und $\sqrt{2} \approx 1.5$, lässt sich der Ausdruck weiter vereinfachen zu 95 % KI = *Differenz* ± *3 x SD*/$\sqrt{n}$
Die Güte dieser Annäherung wurde überprüft, indem die Weite der hiermit berechneten Konfidenzintervalle verglichen wurde mit der Weite von mit der „exakten“ Formel berechneten Konfidenzintervalle, wobei Gruppengrößen zwischen 10 und 100 Probanden und Effektgrößen (hier ausgedrückt als Kehrwert der Standardabweichungen der Differenzen zwischen den Gruppen) von 0,2 bis 0,8 verwendet wurden. Die Annäherung tendierte dazu, Konfidenzintervalle zu erzeugen, die zu weit waren, aber die Fehlergröße lag dabei immer unter 8 %. Dies war sogar der Fall, wenn sich n_t und n_c oder SD_t und SD_c um 20 % unterschieden. Der durchschnittliche absolute Fehler lag bei 5 %. Diese kleinen und zu konservativen Entscheidungen neigenden Fehler sind wahrscheinlich für klinische Entscheidungsfindungen vertretbar.

Kleines Wörterbuch zum Verständnis der Grafiken:
very harmful effect – sehr schädliche Wirkung
smallest clinically worthwhile effect – kleinster klinisch lohnenswerte Effekt
no effect – keine Wirkung
very effective treatment – sehr wirksame Behandlung
treatment is not worthwile – Behandlung lohnt nicht
treatment is worthwhile – Behandlung lohnt sich

nen sie manchmal aus den grafischen Darstellungen abgelesen werden. In manchen Artikeln können die Ergebnisse der Studie auch inadäquat dargestellt sein, sodass es nicht möglich ist, die 95 %-Konfidenzintervalle zu berechnen. Solche Studien sind schwierig zu interpretieren. Manche Artikel über Studien berichten Mediane oder Interquartilabstände (oder andere Maße der zentralen Tendenz bzw. der Streuung) anstelle von arithmetischen Mitteln („Durchschnitten") und SDs. Dann wird es normalerweise nicht möglich sein, Konfidenzintervalle für diese Studien zu berechnen.

Manchmal sind auch Konfidenzintervalle für die Mittelwertdifferenzen wichtig

Die gerade beschriebene Vorgehensweise zur Berechnung von Konfidenzintervallen für Mittelwertdifferenzen zweier Gruppen tendiert dazu, übermäßig konservative Konfidenzintervalle zu erzeugen (d. h. die Konfidenzintervalle sind unter gewissen Umständen zu weit).[10] Dies ist besonders dann der Fall, wenn es sich um eine cross-over-Studie handelt, um eine Studie, bei der die Probanden vor der Randomisierung „gepaart", d. h. paarweise in den beiden Gruppen gegenüber gestellt werden (matched pairs), oder um eine Studie, in der statistische Verfahren angewandt werden, die zum Ziel haben, verschiedene Quellen für die Varianz zu identifizieren (Kovarianzanalysen – analysis of covariance: ANCOVA). Nicht so häufig, nämlich dann, wenn der Stichprobenumfang klein und die Gruppengrößen sehr unterschiedlich sind, kann das Konfidenzintervall auch zu klein ausfallen. In solchen Fällen ist es in hohem Maße wünschenswert, dass die Verfasser auch Konfidenzintervalle für die Mittelwertdifferenzen zwischen den Gruppen berichten. Wenn die Autoren keine Konfidenzintervalle für die Differenzen zwischen den Gruppen berichten, ist es sonst für die Leser meist nicht möglich, präzisere Schätzungen für das 95 %-Konfidenzintervall zu berechnen.

Die nächste Folge dieser Serie wird darstellen, wie man die Größe des Behandlungseffekts bei dichotomen Outcome-Variablen bestimmen kann.

10 Dies deshalb, weil diese Methode nur eine Annäherung mit verminderter Präzision ist. Es entspricht guten wissenschaftlichen Verhaltensregeln, sich dann das „Leben schwerer zu machen", d. h. in diesem Falle, nicht leicht zu eindeutigen – und die ursprüngliche Arbeitshypothese in der Regel bestätigenden – Ergebnissen zu gelangen. Es geht darum, vorschnelle Schlüsse, hier über die Wirksamkeit von Therapie, zu vermeiden; Anm. des Übers.

Literatur

1. Chalmers I, Altman D 1995. Systematic Reviews. London. British Medical Journal
2. Dean CM, Shepherd RB 1997. Task related training improves performance of seated reaching tasks after stroke. A randomized controlled trial. Stroke, 28; 722-728.
3. Dini D, Del Mastro L, Gozza A, Lionetto R, Garrone O, Forno G, Vidili G, Bertelli G, Venturini M 1998. The role of pneumatic compression in the treatment of postmastectomy lymphedema. A randomized phase III study. Annals of Oncology, 9, 187-90.
4. Gardner MJ, Altman DG 1989. Statistics with Confidence – Confidence Intervals and Statistical Guidelines. London. British Medical Journal, 20-33.
5. Guyatt GH, Sackett DL, Cook DJ 1993. User's guide to the medical literature: II. How to use an article about therapy or prevention: Are the results of the study valid? Journal of the American Medical Association, 270; 2598-2601.
6. Guyatt GH, Sackett DL, Cook DJ 1994. User's guide to the medical literature: II. How to use an article about therapy or prevention: What were the results and will they help me in caring for my patients? Journal of the American Medical Association, 271; 59-63.
7. McAlister FA, Straus SE, Guyatt GH, Haynes RB 2000. Users Guide to the medical literature XX. Integrating research evidence with the care of the individual patient. Journal of the American Medical Association, 283; 2829-2836.
8. Sackett DL, Richardson WS, Rosenberg W, Haynes RB 1998. Evidence-based Medicine. How to practice and teach EBM. Edinburgh; Churchill Livingstone; 91-96
9. Sand PK, Richardson DA, Staskin DR, Swift SE, Appell RA, Whitmore KE, Ostergard DR 1995. Pelvic floor electrical stimulation in the treatment of genuine stress incontinence: a multicenter, placebo-controlled trial. American Journal of Obstetrics and Gynecology 173; 72-79
10. Sim J, Reid N 1999. Statistical inference by confidence intervals: issues of interpretation and utilization. Physical Therapy 79; 186-195
11. Zhang J, Bernasko JW, Leybovich E, Fahs M, Hatch MC 1996. Continuous labor support from labor attendant for primiparous women: a meta-analysis. Obstetrics and Gynecology 88; 739-744

Korrespondenzadressen:

Dr. Robert D. Herbert
School of Physiotherapy
The University of Sydney
PO Box 170
Lidcombe NSW 1825, Australia
r.herbert@fhs.usyd.edu.au

Dr. Erwin Scherfer
Bildungswerk Physio-Akademie
des ZVK gGmbH
Wremer Specken 4, 27638 Wremen
E-Mail: e.scherfer@physio-akademie.de

R. D. Herbert

- *Senior Lecturer der Physiotherapieschule an der Universität Sydney, Australien*
- *Unterrichtung von undergraduate und postgraduate Studenten in Evidenz basierter Praxis*
- *Forschungsfelder: Muskelmechanik und klinische Untersuchung der Effekte physiotherapeutischer Interventionen*
- *Direktor des Zentrums für Evidenz basierte Physiotherapie und Wissenschaftlicher Redakteur des australischen Joumal of Physiotherapy*