Methodenkritische Stellungnahme

zum Methodenpapier des Wissenschaftlichen Beirats Psychotherapie nach §11 PsychThG:

„Verfahrensregeln der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie“ – Version 2.9 vom 3.6.2019

Harald Walach (1,2)

1 Change Health Science Institut, Berlin

2 Next Society Institute, Kazimieras Simonavicius University, Vilnius, Litauen

Kontakt

Prof. Dr.Dr. Harald Walach

CHS-Institut

Schönwalder Str. 17

13347 Berlin

030 467 97 436

hw@chs-institute.org

https://harald-walach.de

0 Einführung: Aufgabe, eigene Vorannahmen

Das Methodenpapier des Wissenschaftlichen Beirats Psychotherapie in seiner aktuellen Version 2.9 vom 3.6.2019 definiert die Grundlagen, auf denen ein Antrag eines Psychotherapieverbandes auf Zulassung eines bislang nicht durch das Psychotherapie-Gesetz anerkannten Verfahrens beurteilt wird. Genauer, es legt vor allem fest, wie die Studien, die dem Zulassungsantrag beigelegt werden, beurteilt werden und welche Definitionskriterien angewandt werden. Ich wurde gebeten, es in diesem Gutachten kritisch zu beleuchten.

Aufgabe

Dabei ist meine Aufgabe, aufgrund meiner eigenen Kompetenz und meines Hintergrundes (siehe hierzu *Der Autor* am Ende des Textes) vor allem eine kritische Bewertung der angewandten methodischen Kriterien und ihrer Grundlagen, impliziten Annahmen und Voraussetzungen vorzunehmen. Meine Vorannahmen mache ich im Folgenden im Sinne einer Reflexivität deutlich, die übrigens dem Methodenpapier in erschreckender Weise fehlt.

Reflexivität meiner eigenen Voraussetzungen

Meine Herkunft aus der Tradition der Evaluation komplementärmedizinischer Verfahren (siehe *Der Autors* am Ende des Textes) ist in besonderer Weise der hier anstehenden Frage zuträglich. Denn als jemand, der Psychotherapie zwar aus eigener Praxis kennt, damit aber keinerlei Broterwerb verbindet, habe ich gegenüber dieser Zunft keine Interessenskonflikte. Denn mein Leben würde genauso weitergehen, wenn Psychotherapie im Allgemeinen oder eine spezielle Disziplin verschwände. Andererseits bin ich der Psychotherapie im Allgemeinen und speziell tiefenpsychologischen und erlebenszentrierten Verfahren durch meine Ausbildung wohlgesonnen (siehe hierzu meine Ausführungen in Walach, 2020, orig. 2005). Außerdem bin ich aufgrund meiner langjährigen Erfahrungen mit dem Medizinbetrieb leidenschaftlich der Meinung, dass Medikation für psychische Probleme, wie sie dem herrschenden Trend der Medikalisierung entspricht, die schlechteste Option ist und jede noch so rudimentäre Psychotherapie besser als jedes Medikament ist. Denn sorgfältige Analysen haben ergeben: Die Nebenwirkung von Medikamenten, und dazu gehören vor allem Psychotropika, sind die dritthäufigste Ursache von Todesfällen in der westlichen Welt (Gøtzsche, 2013, 2015). Selbst wenn Gøtzsches Schätzungen übertrieben sind, was ich persönlich aufgrund meiner eigenen Kenntnis nicht glaube, so sind die Nebenwirkungen und schädlichen Folgen psychotroper Medikation weitgehend unterschätzt. Dies hängt auch damit zusammen, dass viele Meinungsführer auf diesem Feld schwerwiegende Interessenskonflikte haben, die nicht immer transparent sind, wie ein Blick auf die Webseite Leitlinienwatch lehrt (https://www.leitlinienwatch.de/, Zugriff 24.1.2023).

Ein von uns kürzlich publizierter Meta-Review, der ein Drittel aller seit 2008 verfügbaren Cochrane-Reviews zufällig einschloss (1.587 Interventionen aus allen Bereichen der Medizin, incl. Psychotherapie, diätetischer und Verhaltensinterventionen) erbrachte, dass nur 5,6% aller Interventionen robuste, klinisch relevante und statistisch signifikante Belege anführen können. Belege für Schaden war bei 8,1% der Reviews zu erkennen, die dieses Outcome dokumentiert hatten. (Howick et al., 2022)

Im Klartext: Wir überschätzen in der Regel die Wirksamkeit medizinischer – vor allem pharmakologischer – Interventionen und vernachlässigen ihr Schadenspotenzial. Das gilt vor allem auch für psychische Erkrankungen. Aus diesem Grunde habe ich eine klare Präferenz für Psychotherapie, egal welcher Couleur.

Meine Methodenreflexion, die auf eine mittlerweile mehr 30-jährige eigene klinische Forschungspraxis zurückgeht, seit ich 1992 als Forschungsmitarbeiter am Psychologischen Institut der Universität Freiburg die Arbeitsgruppe Homöopathieforschung und 1999 am Universitätsklinikum Freiburg die Sektion Evaluation Komplementärmedizin gegründet habe, hat mich ausführlich über die Probleme nachdenken lassen, die dem Methodenpapier des Wissenschaftlichen Beirats zwar inhärent sind, dort aber weder reflektiert noch in erkennbarer Weise aufgegriffen sind.

Dazu gehören u.a. die Problematik eines Versuchs *„spezifische" Effekte* von *„unspezifischen"* zu trennen. Das halte ich für einen kapitalen Fehler jeglicher Evaluationsmethodik. Er ist der Tatsache geschuldet, dass es für die medizinische Zulassungslogik notwendig ist, dies zu tun und von entsprechenden Gesetzen gefordert wird. Dort ist dieser Versuch auch angebracht. Die Übertragung indes auf andere Situation ist, was Whitehead eine „fallacy of misplaced concreteness" (Whitehead, 1978) nennen würde: Eine Fehlübertragung von vermeintlich Konkretem.

Dazu gehört auch der im Methodenpapier zwar angesprochene, aber nicht durchdachte *Konflikt zwischen externer und interner Validität*.

Ein weiteres Element dieses Problembündels ist die unreflektierte Bevorzugung der *Randomisation*. Diese ist als methodisches Element zwar theoretisch gut gesichert, aber empirisch schlecht belegt. Überdies gibt es für kleine Studien bessere Methoden, die von dem Methodenpapier gar nicht berücksichtigt werden

Ich werde diese drei Punkte separat verhandeln und in einem vierten Punkt auf konkrete Widersprüchlichkeiten und Inkongruenzen im Methodenpapier eingehen.

Ich werde bewusst keine inhaltlichen Qualifikationen zu Elementen psychotherapeutischer Verfahren vornehmen, wie dies Bruce Wampold in seinem Gutachten (Wampold, 2021) bereits unternommen hat und habe bezüglich der Überlegenheit, Unterlegenheit oder Zulassungswürdigkeit bestimmter Verfahren weder eine dezidierte Meinung noch persönliches Bedürfnis. Ich werde allenfalls die Problematik der wissenschaftlichen Belege gestalttherapeutischer und emotional-prozessbetonter therapeutischer Arbeit am Ende berühren.

1. Der Synergismus zwischen „spezifischen" und „unspezifischen" Effekten, das Wirksamkeitsparadox und das vergebliche Suchen nach dem „Gold des Spezifischen"

Unspezifische und Spezifische Therapie-Effekte – eine kurze Geschichte

Der Versuch, sog. „spezifische" von „unspezifischen" oder intendierte von unintendierten Effekten zu trennen entstammt dem medizinischen Versuch, pharmakologische Wirksamkeit nachzuweisen, wie sie etwa seit dem Siegeszug der modernen Pharmakologie als notwendig erachtet wurde. Der deutsche Pharmakologe Martini (Martini, 1932) hat noch vor dem Krieg als einer der ersten in seiner Methodenlehre Placebokontrollen gefordert, um „wahre" von „falschen" oder „eingebildeten" Effekten trennen zu können. Als durch die Erfahrungen mit Placebo als Ersatz für fehlendes Morphium im zweiten Weltkrieg Beecher die medizinische Öffentlichkeit auf die Bedeutung des Placebo-Effektes hingewiesen hatte (Keats &

Beecher, 1950), griffen verschiedene Methodiker in den Conferences on Therapy diese Einsicht nach dem Krieg auf und forderten Placebo-Kontrollen (Conferences on Therapy, 1946, 1954).

Für eine *Regulation pharmakologischer Substanzen*, und aus meiner Sicht nur für diese, ist die Logik der Trennung spezifischer und unspezifischer Effekte überzeugend. Um psychologische Effekte auszuschließen, wurde die doppelte Verblindung eingeführt, deren Vorausseztung die Anwendung eines *ununterscheidbaren* Placebos ist (Kaptchuk, 1998, 2001). Denn nur so lassen sich psychologische von pharmakologischen Effekten trennen.

Fehlschluss Nummer 1: Es ist ein Fehlschluss pharmakologische Denkmodelle auf psychologisch-psychotherapeutische Logik anzuwenden.

Denn bei psychotherapeutischen Methoden geht es ja gerade darum, therapeutische Effekte mit Hilfe und unter Maximierung unserer Kenntnisse von Psychologie zu erzeugen, und nicht, sie auszuschließen. Zwar werden in der Psychotherapie sehr unterschiedliche Methoden und Verfahren angewandt. Aber sie alle sind psychologischer Natur, und es wird in der Psychotherapieforschung kein „Substanzeffekt“ von einem „nur psychologischen“ Effekt unterschieden, sondern allenfalls, etwa im Rahmen einer Prozessforschung, die differentiellen Effekte unterschiedlicher psychologischer Prozesse und Interventionen. Dennoch hat in letzter Zeit gerade auch in der von der Verhaltenstherapie und von der Psychiatrie beeinflussten Forschung das Denken an Einfluß gewonnen, das psychotherapeutische Verfahren wie ein Spezifikum betrachtet, das man beliebig abfüllen und kontextfrei testen können soll. (Rief & Gaab, 2016) Dies wurde schon vor Zeiten scharfsichtig kritisiert. (Kriz, 2000)

Insofern ist die Übertragung einer pharmakologischen Zulassungslogik auf Psychotherapie ein medizinischer Fehlschluss.

Die Unbrauchbarkeit eines additiven Komponentenmodells von Therapie-Effekten und das Wirksamkeitsparadox

Die wissenschaftliche Effektivitätstestung geht von einem additiven Komponentenmodell aus. Therapie-Effekte, so die Logik, setzen sich aus unterschiedlichen Komponenten zusammen, die unterschiedlich stark erwünscht und theoretisch unterschiedlich starke Gewichtung haben. Dazu gehören in der Psychotherapie neben den erwünschten und postulierten spezifischen Interventionseffekten der Methode auch Meßfehler, wie etwa die statistische Regression zur Mitte bei Wiederholungsmessungen oder der natürliche Zeitverlauf einer Erkrankung. Dazu gehören aber auch unspezifische psychologische Effekte einer Behandlung. Bei einer pharmakologischen Testung wären das etwa Erwartung, Hoffnung, Beruhigung und Entspannung, Beziehungsmöglichkeit, Konditionierung. Bei einer psychotherapeutischen Studie wären das die sog. allgemeinen Wirkfaktoren („common factors“im Sinne von Frank (1961) der Beziehung (Hartmann-Kottek, 2021), der Ressourcenaktivierung, dem Anbieten eines guten Erklärungsmodelles, und noch mancher anderer (Grawe, 1998, 1999; Grawe & Grawe-Gerber, 1999).

Bei einer Medikamententestung kommen noch spezifische pharmakologische Effekte hinzu. Diese allein gelten den Zulassungsbehörden als wertvoll und ihre Überlegenheit über alle unspezifischen Faktoren soll abgesichert werden. Analog versuchen viele Psychotherapieforscher, verfahrensspezifische Effekte forscherisch zu sichern und nennen diese dann „spezifische Effektivität“. Dieses Denken ist auch dem Methodenpapier inhärent. Dieser

Zusammenhang ist verkürzt und schematisch in einem Gedankenexperiment dargestellt (Abbildung 1).

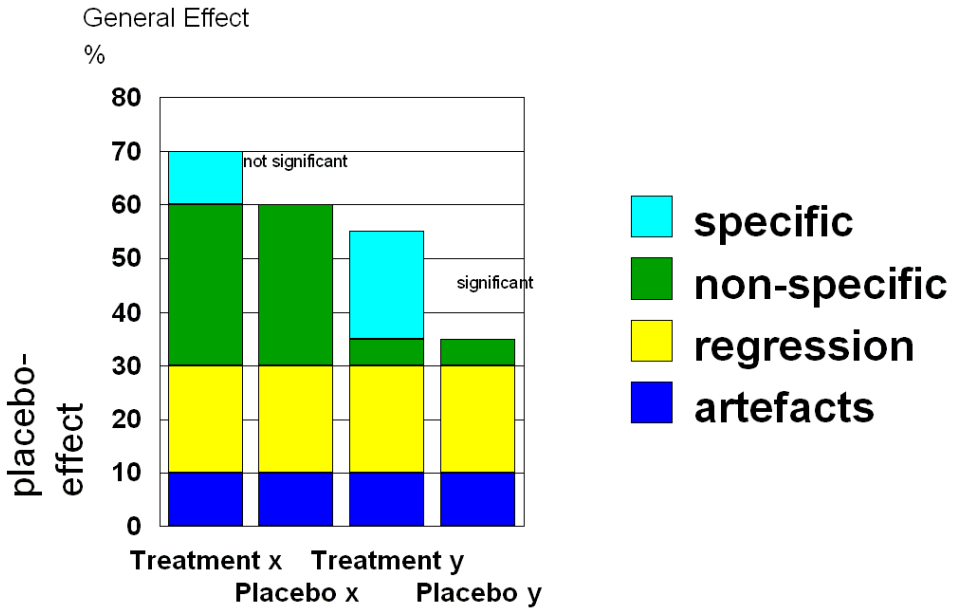

Abbildung 1 - Veranschaulichung des additiven Komponentenmodells und des Wirksamkeitsparadoxes

Abbildung 1 illustriert ein Gedankenexperiment, das „Wirksamkeitsparadox". Es geht von zwei unterschiedlichen Interventionen bei der gleichen Grundkrankheit bei ähnlichen Patientengruppen aus, sagen wir Gestalttherapie bei Depression (Behandlung x) und pharmakologische Therapie bei Depression (Behandlung y), die beide in je einer kontrollierten Untersuchung untersucht worden sind. Bei der pharmakologischen Therapie wurde gegen Placebo getestet, bei Gestalttherapie (Therapie x) sei der Vergleich eine vermutete schwächere Placebotherapie ohne die vermeintlich „spezifische" Ingredienz. Die Logik der Subtraktion von Therapie-Effekten zeigt, so unsere Annahme, dass die konventionelle Behandlung, Behandlung y, erfolgreich, weil signifikant ist. Denn die spezifischen Effekte sind hier grösser und daher auch statistisch signifikant (bei adäquater statistischer Mächtigkeit), während sie bei Behandlung x nicht statistisch signifikant sind (bei ähnlicher statistischer Mächtigkeit).

Man sieht sofort, dass das subtraktive Modell intuitiv falsch ist. Dies ist das Wirksamkeitsparadox. (Walach, 2001, 2016) Es kann nämlich eine Behandlungsmethode, die es schwer hat, ihre Spezifität nachzuweisen, hier Behandlung x, insgesamt und global wesentlich effektiver sein als eine andere Methode, die zwar als „wirksam" belegt ist, aufgrund der hier geltenden subtraktiven Logik (Behandlung y). Dennoch kann diese erwiesen „wirksame" Methode wesentlich schlechtere therapeutische Effekte erzielen.

Dies liegt daran, wie man bei einem Blick auf Abbildung 1 sofort sieht, dass die sog. „unspezifischen" Effekte keine feste Größe darstellen, sondern modifizierbar sind. Hinter ihnen verbirgt sich ein ganzes Arsenal möglicher psychologischer Effekte, von der Konditionierung, über die Hoffnung und Erwartung, bis hin zu unspezifischen Entspannungen und den in der Psychotherapie wichtigen Effekten einer guten Beziehungsgestaltung, der Aktivierung von Ressourcen und vieles andere mehr.

Ich habe das Wirksamkeitsparadox 2001 in die Diskussion gebracht, weil ich illustrieren wollte, wie kurzsichtig die Fixierung auf vermeintliche „Spezifität“ in Kontexten ausserhalb pharmakologischer Regulation, aber eigentlich auch dort, ist. Damals war das ein Gedankenexperiment. Mittlerweile hat die Empirie die tatsächliche Gültigkeit dieses Gedankenexperiments bestätigt: Unterschiedliche Placebos haben in der Migränetherapie unterschiedliche Mächtigkeit und zwar dergestalt, dass manche Effekte von Placebos die Effekte spezifischer Therapie übertreffen (Meissner et al., 2013); kognitive Verhaltenstherapie ist in dieser Netzwerkanalyse einem Akupunkturplacebo nicht überlegen. Sollte man deshalb fordern, dass kognitive Verhaltenstherapie bei Migräne durch Akupunkturplacebo ersetzt wird, weil letzteres billiger ist? Die großen deutschen Akupunkturstudien (GERAC Studien) zeigten, dass Placebo-Akupunktur in allen drei getesteten Modalitäten (Migräneprävention, chronische Rückenschmerzen, Arthroseschmerzen) dem besten, was Evidence-Based Medicine in Deutschland zu bieten hat (und was selber ja gegen unspezifische Wirksamkeit vorgibt abgesichert zu sein) etwa um den Faktor 2 überlegen oder gleichgut ist (Diener et al., 2006; Haake et al., 2007; Scharf et al., 2006).

Wir halten fest: vermeintlich unspezifische Therapie-Effekte können wirksamer oder gleich wirksam sein als klar abgesicherte spezifische Therapien. Das ist in der Medizin mittlerweile belegt und dürfte für die Psychotherapieforschung nicht anders sein, wie Wampolds Arbeiten belegen (Wampold, 2021; Wampold, Frost, & Yulish, 2016; Wampold & Imel, 2015). Das ist deswegen so, weil die Annahmen, die gemacht werden müssen, damit das additive oder subtraktive Modell funktioniert, falsch sind.

1. Die Effekte sind höchstwahrscheinlich nicht additiv, sondern multiplikativ und damit synergistisch. Mein Beispiel für ein synergistisches System (Walach, 2011) ist ein kleines Kind, das gut reiten kann. Es kann auf seinem Pferd schneller unterwegs sein und höher springen als es das alleine könnte und als es das Pferd von sich aus täte, ausser es wäre in Gefahr. Daher ist der Versuch, durch ein subtraktives Design auf spezifische Effeke zu schließen, von vorneherein zum Scheitern verurteilt. Dies gilt vermutlich sogar für pharmakologische Verfahren, aber mit großer Sicherheit für komplexe Therapieverfahren wie für die Psychotherapie.
2. Das additive Modell nimmt an, dass Placebo-Effekte oder unspezifische Effekte Fehlervarianz darstellen, die sich über die Studien hinweg ausmittelt (McQuay, Carroll, & Moore, 1996). Genau das ist aber nicht der Fall. Unsere eigene Meta-Analyse von 144 pharmakologischen Langzeitstudien zeigt: die Korrelation zwischen Verum-Behandlungserfolg und Placebo-Behandlungserfolg ist r = .78 und nicht durch Eigenschaften der Studien oder durch die behandelte Diagnose zu erklären (Walach, Sadaghiani, Dehm, & Bierman, 2005). Wären Placebo-Effekte Fehler Varianz, müsste die Korrelation gleich oder nahe Null sein. Wir haben diesen Befund soeben in einer neuen Analyse, die wir zur Publikation vorbereiten, erhärtet.
3. Es gibt kein inertes Placebo, nicht einmal in der Pharmakologie. Der Versuch ein „Psychotherapie-Placebo“ zu finden ist einem obsoleten Denk- und Forschungsmodell geschuldet, das ausser der Tradition keinerlei vernünftiges Argument, geschweige denn eine solide Empirie zur Stützung heranziehen kann. Expertenkonsens ohne Daten ist aber genau das, was die Bewegung der „Evidence Based Medicine“ versucht hat abzuschaffen.

Fehlschluss Nummer 2: Es ist irrig zu glauben, man könne ein inertes Psychotherapie-Placebo entwickeln und dessen Effekte von einer wahren Behandlungsgruppe subtrahieren, um einen „echten“ Therapie-Effekt zu isolieren.

Denn die sog. „spezifischen“ Effekte sind wie Zwerge, die auf den Schultern von Riesen sitzen, wie ich einmal das bekannte Fenstermotiv von Chartres uminterpretiert habe (Klibansky, 1936). Die Zwerge sind nur deswegen weitsichtig, weil sie auf den Schultern von Riesen, den unspezifischen Effekten sitzen, und die Riesen gewinnen ihre Funktion nur daraus, dass sie die Zwerge tragen. Auch dieser Vergleich hinkt sehr. Denn im Grunde soll er sagen: Es ist unsinnig ein synergistisches System zu zerlegen. Man erhält dadurch keine wirksamen Komponenten, sondern lediglich sinnlose und zusammenhangslose Teile.

2. Die Inkompatibilität von Interner und Externer Validität

Das Methodenpapier hat richtig erkannt: Interne und externe Validität sind kaum in einer Studie gemeinsam zu maximieren. Es behauptet zwar, das ginge, aber ohne dafür einen Beleg anzuführen. Mir ist in meiner über 30-jährigen Erfahrung und nach der Lektüre von vielen tausend klinischen Studien, aus der Medizin und aus der Psychotherapie, keine einzige Studie bekannt, die diese Quadratur des Kreises schafft.

Daraus habe ich die konzeptuelle Konsequenz gezogen und behaupte: *Interne und externe Validität sind inkompatible Konzepte*, inkompatibel im gleichen Sinne wie nicht-kommutierende Observable in der Quantenmechanik inkompatibel sind (Atmanspacher, Römer, & Walach, 2002; Walach & Loef, 2015). Das bedeutet: Man kann nicht die eine Grösse verwenden, um die andere zu definieren. Vielmehr sind sie in unterschiedlichen Koordinatensystemen zuhause, so ähnlich wie man eine y-Achse in einem kartesischen System nicht durch die x-Achse ausdrücken kann. Projiziert man die y-Achse auf die x-Achse, ist der Projektionspunkt Null, und umgekehrt.

Daher müssen interne und externe Validität in unterschiedlichen Studientypen untersucht werden, idealerweise in Ergänzung zueinander. Dies hat auch das Methodenpapier intuitiv erkannt, indem es zum Beispiel zwei sich selbst widerprechende Kriterien verwendet:

Das wichtige Kriterium B.8, das Randomisation oder Parallelisierung (oder Quasi-Randomisation) fordert, und zwar als Ausschlusskriterium, widerspricht dem Kriterium C.8, das u.a. die externe Validität definiert, wo es heisst: 1. Patienten entscheiden sich selbst; 2. Teile der Patienten weden zufällig zugewiesen (wobei „1“ jeweils die beste Note darstellt).

Dies ist ein sehr typischer und für die Psychotherapieforschung – wie für jede Therapieforschung – zentraler Konflikt. Denn vielleicht das wichtigste Element einer jeden Therapie ist die Tatsache, dass ein Patient oder Klient Verantwortung für seine Situation übernimmt und aktiv wird. (Walach & Loughlin, 2018) Das bedeutet: Er oder sie entschließt sich eine Therapie aufzusuchen. Diese Aktivität wird aber konterkariert durch eine Zufallszuweisung. Entweder ist jemand aktiv und wählt und ist dann unwillig, in einem Randomisationsverfahren den Zufall über seine Zukunft entscheiden zu lassen. Oder jemand ist inaktiv. Insofern hat der Konflikt zwischen Kriterium B.8, Randomisation, und C.8, Patienten dürfen wählen, eine Basis in der Wirklichkeit und kann nicht einfach übergangen werden. Außerdem sind Patienten, die sich randomisieren lassen nicht mit denen vergleichbar, die das nicht tun, wie Wampold in seinem Gutachten ebenfalls erläutert. Daher beeinträchtigt Randomisation, und damit die Erhöhung der internen Validität, ipso facto die externe Validität.

Deswegen ist es sowohl sachlich als auch konzeptionell-therapeutisch falsch, randomisierte Studien überzubewerten (B.8 Ausschlusskriterium; hingegen C.8 eines von mehreren Kriterien). Seligman hat vor Zeiten die entsprechende Konsequenz gezogen und in seiner Consumer Study an einer sehr großen Zahl von Patienten Outcomeforschung betrieben, die robuste, extern valide Daten vorlegte. (Seligman, 1995) Es hat seinem Ruf nicht geschadet, wie Kriz (2000) feststellte. Drei Jahre später wurde er Vorsitzender der APA. Der Wissenschaftliche Beirat lässt in seinem Methodenpapier eine Anerkennung dieser Sichtweise vermissen.

Es wäre angemessen, diesen Sachverhalt zur Kenntnis zu nehmen. Die Lösung, die wir vorgeschlagen haben lautet: statt einem hierarchischen Studienmodell, bei dem fast ausschließlich interne Validität bewertet wird und externe Validität allenfalls ein Zusatzkriterium darstellt, sollte man ein zirkuläres Modell ins Auge fassen (Walach, Falkenberg, Fonnebo, Lewith, & Jonas, 2006). Dieses geht davon aus, dass es nicht „die" beste Methodik gibt, sondern nur angemessene Methoden zur Beantwortung von unterschiedlichen Fragestellungen. Diese Methoden ergänzen sich und stehen nicht in einem Abhängigkeits- oder Unterordnungsverhältnis. Das bedeutet konkret: Während eine randomisierte Studie, die aktiv kontrolliert ist, etwas über eine mögliche Überlegenheit oder Gleichwertigkeit gegenüber einer Konkurrenzbehandlung bei Patienten aussagen kann, die selber keine Präferenz haben, sagen größere Kohortenstudien ohne Randomisation etwas darüber aus, wie gut Verfahren bei den Patienten wirken, die sich dafür entscheiden und bei lebensechten Therapeuten in der freien Praxis behandelt werden. Diese beiden Elemente von Information ergänzen sich, und stehen nicht miteinander in Konkurrenz. Man könnte, wie wir praktisch gezeigt haben, einen solchen Ansatz auch im Rahmen eines Bayesianischen Ansatzes zu einer Meta-Analyse über unterschiedliche Studientypen nutzen (Klement, Bandyopadhyay, Champ, & Walach, 2018), aber das nur am Rande.

In jedem Falle wären zum Beispiel sorgfältige und ausreichend große zwei- oder mehrarmige Kohortenstudien in der niedergelassenen Praxis, die Patienten und ihre Therapie-Erfolge oder -misserfolge in den von ihnen gewählten Therapien über längere Zeiträume dokumentieren mindestens genauso aussagekräftig wie randomisierte Studien, die dem Kriterium B.8 genügen würden und vielleicht in allen anderen Kriterien des Komplexes B sehr gute Bewertungen erhalten, also intern sehr valide sind, aber mit dem Kriterium der externen Validität gerade noch akzeptable Werte erreichen.

Das klassische Beispiel für diese Situation sind Studien zu Antidepressiva. Auch wenn ihre gemeinsame Effektstärke klinisch gering ist (etwa d = 0.38 gegenüber Placebo) (Kirsch et al., 2008; Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008), so gelten sie doch als „wirksam". Sie erreichen diese Wirksamkeit nur, indem Zulassungsstudien im Sinne einer Homogenisierung Einschlusskriterien und Ausschlusskriterien definieren, die in der Regel nur solche Patienten in Studien aufnehmen, die keine multiplen Diagnosen ausser Depression aufweisen. Meistens werden ihnen auch vor Aufnahme in die Studie andere Substanzen entzogen, was ihre Empfänglichkeit für eine neue Substanz gegenüber der Placebogruppe erhöht und damit auch eine Unterschiedsmaximierung bewirkt (Gøtzsche, 2015). In der Praxis aber kommen solche „reinen" Depressionspatienten kaum vor. Die meisten haben auch eine Angststörung, Persönlichkeitsstörung oder Abhängigkeitsstörung (und wenn es „nur" die Abhängigkeit von Antidepressiva ist). Daher sind die „wirksamen" Antidepressiva in der Praxis häufig unbrauchbar, wie sich in der STAR*D-Studie, vor allem in ihrer Re-Analyse, eindrucksvoll gezeigt hat (Kirsch, Huedo-Medina, Pigott, & Johnson, 2018; Pigott, Leventhal, Alter, & Boren, 2010; Rush et al., 2006; Trivedi, Fava, et al., 2006; Trivedi, Rush, et al., 2006).

Fehlschluss Nummer 3: Es ist sachlich falsch, interne Validität gegen externe Validität auszuspielen und höher zu bewerten. Das hierarchische Forschungsmodell ist sachlich nicht gerechtfertigt und sollte durch ein zirkuläres abgelöst werden.

3. Die Überbewertung und sachlich falsche Handhabung der Randomisation

Randomisation ist ein Wundermittel, so scheint es. Man weist Patienten per Zufall auf Gruppen zu und hat im Handumdrehen die Hauptgefahr kausaler Schlussfolgerung gebannt: den Einfluss konfundierender Faktoren. So lautet die Theorie. Allerdings ist es wichtig zu verstehen: das ist Theorie. Keine schlechte Theorie, aber Theorie immerhin. Es gibt wenig empirische Belege, die diese Theorie tatsächlich empirisch untermauern. Dort wo die Theorie, zum Beispiel durch Simulationsstudien, untersucht wurde, zeigt sich: erst bei sehr großen Studiengruppen von ca. 300 Personen, greift der Zufall wirklich so zu, dass sich konfundierende Variablen zufällig verteilen (Aickin, 1983, 2001, 2002). Das Methodenpapier scheint das zumindest implizit zu erkennen, indem es bei Kriterium B.8 und A.5 eine Stichprobengröße von > 30 fordert. Allerdings ist diese Zahl willkürlich und viel zu gering, um auch nur annähernd zu garantieren, dass die Randomisation eine Homogenisierung konfundierender Variablen erreicht. Warum nicht n > 25 oder n > 45? Es gibt kein einziges vernünftiges Argument, das „30“ zu einer magischen Zahl macht, außer dem statistischen, dass ab 30 Freiheitsgraden die t-Verteilung in die Standardnormalverteilung übergeht. Aber das hat nichts mit dem praktischen Erfolg einer Randomisation zu tun.

Andererseits ist schon aus praktischen Erwägungen eine Teilnehmerzahl von 300 Patienten nur selten zu erreichen, weil der logistische und finanzielle Aufwand sehr hoch wäre.

Kleine randomisierte Studien greifen jedoch oft zu Kunstgriffen, wie Blockrandomisation oder Stratifizierungen. Blockrandomisation schränkt den Zufall ein mit dem Ziel, Gruppengrößen gleich zu halten. Das wiederum geschieht, weil die kleinste Gruppe die Power der gesamten Studie bestimmt. Also versucht man, eine Reduktion der statistischen Mächtigkeit zu vermeiden. Durch die Einschränkung des Zufalls fällt aber ein entscheidender Vorteil der Randomisation weg, nämlich die wirklich zufällige Verteilung von Personen zu Gruppen.

Das Kriterium B.9, „Vergleichbarkeit der Gruppen“, lässt erkennen, dass, wer auch immer diesen Abschnitt geschrieben hat, die grundlegende Logik der Randomisation nicht verstanden hat. Denn hier wird von „signifikanten“ Unterschieden gesprochen, die nicht vorhanden sein dürfen. Das lässt darauf schließen, dass statistische Tests über die Verteilung der Baseline-Variablen angedacht sind. Genau das wird aber von neueren Regeln zur Berichtlegung über randomisierte klinische Studien, etwa dem CONSORT-Statement (Moher et al., 2010; Schulz, Altman, Moher, & CONSORT Group, 2010; Zwarenstein et al., 2008), explizit als unbrauchbar angeführt. Denn entweder geht man von der Theorie der Randomisation aus. Dann wird das Ergebnis keinem statistischen Test unterzogen, weil ja der Zufall selbst, die Basis statistischen Testens, am Werk ist. Oder man testet, weil man dem Zufall nicht traut. Dann ist die Randomisation in ihrer Konzeption nichtig.

Die Debatte zwischen der Nützlichkeit randomisierter Studien im Vergleich zu nicht-randomisierten Kohortenstudien hat auch in der Medizin eine lange Tradition (Black, 1996). Daten zeigen, dass sich vor allem bei großen Kohortenstudien die Ergebnisse von

randomisierten Studien nicht unterscheiden (Concato, 2012; Concato & Horwitz, 2004; Concato et al., 2010; Concato, Shah, & Horwitz, 2000; Koch, Hörmann, Löwel, & Senges, 1998; Linde, Scholz, Melchart, & Willich, 2002; Miller & Joffe, 2011). Das bedeutet: es ist eigentlich wichtiger, große Studien, als randomisierte Studien durchzuführen. Aber wenn man schon, aus welchen Gründen auch immer, nur relativ kleine Studien durchführen kann, und damit meine ich Studien die deutlich kleiner als n = 100 sind, dann sollte man besser Minimierungs- oder Minimalisierungsalgorithmen anwenden statt Randomisation, weil diese mit größerer Wahrscheinlichkeit vergleichbare Gruppen erzeugen (Aickin, 2001, 2002). In einem solchen Prozedere werden einige wichtige konfundierende Variablen, z.B. Alter, Geschlecht, Schwere der Erkrankung, etc. im Vorhinein definiert und erfasst. Die ersten Patienten werden zufällig oder alternierend auf Gruppen zugewiesen. Nach einer gewissen Anzahl, z.B. 10 Patienten pro Gruppe, wird ein regressionsanalytischer Algorithmus aufgebaut, der über eine logistische Regression feststellt, in welche Gruppe der nächste Patient mit einem bestimmten Merkmalscluster geschickt werden muss, um Gleichverteilung wichtiger Variablen zu gewährleisten.

Auch dies ist kein Verfahren, das Patienten Wahlfreiheit zusichert, aber es ist mindestens eines, das bei kleinen Gruppen bessere Chancen auf Vergleichbarkeit der Gruppen bietet. Die Tatsache, dass dieses Verfahren nicht benannt wird, lässt einiges an Methodenkenntnis der Autoren dieses Papiers vermissen.

Randomisation ist kein Gütekriterium an sich, sondern kann nur im Zusammenhang mit anderen Kriterien Auskunft über interne Validität geben. Interne Validität über externe Validität zu stellen, wie dies das Punkteschema des Methodenpapiers erkennen lässt, ist sachlich nicht gerechtfertigt.

Fehlschluss Nummer 4: Es ist sachlich nicht gerechtfertigt, die fehlende Randomisation zu einem Ausschlusskriterium für Studien zu machen, die für eine Bewertung von Psychotherapieverfahren herangezogen werden können.

4. Inkonsistenzen im Bewertungsschema

Die drei oben ausgeführten Punkte sind aus meiner Sicht die wichtigsten Punkte, die gegen eine Anwendbarkeit dieses Schemas zur Bewertung von Psychotherapie-Studien sprechen. Abgesehen davon enthält das Schema Inkonsistenzen und Widersprüche. Ich erwähne sie nicht alle, sondern nur die wichtigsten.

Die Bewertung von interner und externer Validität widersprechen sich

Den Widerspruch zwischen Kriterium B.8 und C.8 habe ich schon erwähnt.

Das Prinzip der reichhaltigen Messung und das Prinzip der Definition eines Hauptzielkriteriums widersprechen sich teilweise

Der Kriterienkatalog sieht als methodisches Gütekriterium die Definition eines primären Zielkriteriums vor (A.7.1). Sind nur die Kriterien genannt, aber nicht definiert (A.7.2), werden 2 Punkte vergeben. Das ist wichtig, weil der durchschnittliche Wert aller Kriterien aus dem Katalog A, sowie aus anderen 2.25 nicht unterschreiten darf. D.h. ein schlechterer Wert auf diesem oder anderen Items reduziert die Fehlertoleranz in anderen

Bereichen. Item A.10, multiple Informationsquellen widerspricht diesem Item direkt. Multiple Informationsquellen im Sinne einer Mehrebenenmessung und einer multiperspektivischen Abbildung von Effekten wird in der psychologischen Evaluationsforschung seit den 80er Jahren als Standard gesehen (Wittmann, 1985, 1988). Daraus hat Grawe u.a. das Prinzip der „reichhaltigen Messung" abgeleitet (Grawe, Donati, & Bernauer, 1994), das seither zum Standard-Arsenal psychologischer Evaluationsmethodik gehört. Und das ist gut so. Denn es bedeutet, dass Effekte sich auf unterschiedlichen Ebenen zeigen und von unterschiedlichen Perspektiven unterschiedlich wahrgenommen werden können. Und genau das soll abgebildet werden.

Die medizinische Evaluationsforschung hingegen ist wesentlich stärker einem positivistischen Entscheidungsmodell verpflichtet, in dem eine simple Entscheidung anhand eines einzigen Kriteriums gefällt werden kann, z.B. lebendig oder tot, gesund oder krank. Dies ist, in Grenzen, im Rahmen der Medizin auch sinnvoll. Aber die Übertragung dieses Entscheidungsmodells auf die Evaluation psychotherapeutischer Verfahren, wie sie in der Definition eines „Hauptzielkriteriums" zum Ausdruck kommt, ist ein direkter Widerspruch zum Mandat einer reichhaltigen Forschung.

Widerspruch zwischen Forschungsethik und Rekrutierungswegen

Die Logik der Bewertung, die diesem Methodenpapier zugrunde liegt, führt zu einem ethischen Dilemma. Die Ethik der Randomisation setzt „equipoise", also Ausgewogenheit zwischen Wissen und Unwissen, voraus, also eine Unentschiedenheit zwischen Alternativen, weil wir nicht genau wissen, was besser ist, und zwar nicht für uns, sondern für die Patienten (Black, 1996; Djulbegovic et al., 2012; Lilford & Jackson, 1995; London, 2017; Miller & Joffe, 2011). Das ist die ethische Voraussetzung dafür, dass überhaupt Zufallszuweisungen ethisch zulässig sind. (Lilford & Jackson, 1995) Sobald eine wie auch immer geartete Kenntnis über mögliche Wirkungen vorliegen, ist diese Unentschiedenheit nicht mehr vorhanden. Mindestens bei qualifizierten Therapeuten, die im Rahmen zum Beispiel einer Universitätsambulanz, einer niedergelassenen Praxis oder einer ähnlichen Einrichtung arbeiten, dürfte das bei etablierten Therapieverfahren nicht (mehr) der Fall sein. Vielmehr haben diese Menschen ihre fachlich-ideologische Zugehörigkeit zu bestimmten Therapieverfahren. Wenn Sie nun – Kriterium C.2 – Patienten über gängige Zugangswege wie Überweisung, Selbstselektion der Patienten, die vom guten Ruf der Einrichtung gehört haben zugewiesen erhalten, dann ist bereits eine implizite Vorentscheidung gefallen. Solche Patienten dann im Rahmen eines Randomisationsverfahrens wiederum in die Unsicherheit zu schicken, „weil wir nicht genau wissen, was besser ist", ist nicht nur inhaltlich schwer vermittelbar. Es scheint mir auch ein ethisch schwer auflösbares Dilemma zu sein. Würde man Patienten über eine Annoncenkampagne rekrutieren, wie man sie häufig in der Berliner U-Bahn sieht („Sie leiden an Depression und haben schon Vieles ausprobiert? Wir haben eine neuartige und interessante Therapiemethode, die wir untersuchen, von der wir denken Sie könnte Ihnen helfen. Melden Sie sich bei…"), dann wäre das ethisch unproblematisch. Denn solche Patienten wissen in der Tat nicht, was sie tun sollen und ob überhaupt etwas geschehen soll. Aber genau dieses Rekrutierungsverfahren schließt Item C.2 aus, genauer gesagt bestraft es mit einer drei.

Widerspruch zwischen Manualisierung (Erhöhung der internen Validität) und unveränderter Praxis (externe Validität)

Dass interne und externe Validität konzeptuell inkompatibel sind, habe ich oben erläutert. Man sieht dies auch am Widerspruch zwischen Item C.4 (Externe Validität: „Intervention wie in klinischer Praxis“ [1], oder „nur leicht verändert“ [2] und Item B.3 (Interne Validität: „Operationale Definition der Intervention, durch ein Therapiemanual“ [1] oder „klare Beschreibung“ [2]). Die wenigsten Therapeuten werden in der Praxis ein einfaches Manual anwenden, sondern ihre Intervention auf die spezielle Situation eines Patienten zuschneiden. Manualisierung und Therapie wie im Praxisalltag widersprechen sich, außer vielleicht in sehr spezifischen Fällen, wenn ein Therapeut eine sehr fokussierte Therapie bei einem ganz eng umgrenzten Syndrom anwendet, aber das dürfte die Ausnahme sein. Wie Wampold (2021) in seinem Gutachten festhält: die empirische Situation hat deutlich gemacht, dass Therapeuten in der Anwendung im Lauf der Zeit und mit zunehmender Erfahrung von Manualen wegkommen. Unsere eigene Befragung an einer repräsentativen Stichprobe von deutschen Psychotherapeuten hat belegt, dass sich die wenigsten einer einzigen Schule zugehörig definieren, auch wenn sie für kassenrechtliche Abrechnungszwecke so definiert sind (Hofmann & Walach, 2011).

Der Kriterienkatalog benachteiligt wichtige Designs, z.B. Wartegruppendesigns

Item B.5 (Strukturelle Äquivalenz bei Kontrollbedingung) benachteiligt wichtige Designs, wie etwa ein Wartegruppendesign oder pragmatische randomisierte Studien. Bei einem Wartegruppendesign müssen Mitglieder der Kontrollgruppe eine gewisse Zeit lang warten, bis sie die Therapie erhalten. Bei einer pragmatischen randomisierten Studie werden zwei völlig unterschiedliche therapeutische Modalitäten verglichen. Jedes dieser Designs, obwohl durchaus valide, erfüllt die Stufen 1 und 2 dieses Items nicht (strukturelle Äquivalenz hinsichtlich Zuwendung). Bei einem Wartegruppendesign gibt es keine Zuwendung. Bei einer pragmatischen randomisierten Vergleichsstudie kann es sein, dass eine Psychotherapie zum Beispiel mit einem gesponsorten Ferienprogramm verglichen wird, bei dem es keinerlei Zuwendung gibt, aber möglicherweise ein anderes aktives Prinzip, von dem man annimmt, es könnte heilsam sein. Das Mayo-Klinik-Konsortium hat vor Zeiten eine solche Studie durchgeführt, bei dem Ärzte entweder an einem strukturierten Training zur Prävention von Burnout teilgenommen haben, oder einfach eine Stunde pro Woche frei bekamen (West et al., 2014). „Strukturelle Äquivalenz der Zuwendung“ ist dabei nicht erkennbar, und doch ist das Ergebnis sehr aussagekräftig. Solche Designs würden benachteiligt werden, obwohl sie möglicherweise klug und valide sind.

Die allgemein dominante und favorisierte kognitive Verhaltenstherapie hat ihre Wirksamkeit anfangs praktisch nur über einfache AB-Einzelfalldesigns und Wartegruppenstudien belegt. Ich halte es für nicht vertretbar, dass Kriterienkataloge verabschiedet werden, die für spätere Aspiranten zur Aufnahme in den Kanon der approbierten Psychotherapien nun plötzlich verschärfte Bedingungen vorsehen, bzw. dass potente und aussagekräftige Designs damit a-priori benachteiligt werden.

Wartegruppenkontrollen sind keine starken Kontrollen, aber sie sind ökologisch sehr valide. Denn man kann Patienten auch bei Therapien, die bereits Wirksamkeitsbelege vorgelegt haben, vermitteln, dass nicht alle Patienten sofort behandelt werden können. Manche werden aus Kapazitätsgründen warten müssen. In einer solchen Situation ist es am einfachsten und gerechtesten, auszulosen wer warten muss. Solche Designs kontrollieren auf jeden Fall eine spontane Veränderung durch den natürlichen Verlauf und damit auch die Regression zur Mitte. Dies sind die wichtigsten konfundierenden Faktoren, die die interne Validität bedrohen.

Pragmatische Kontrollstudien können Auskunft über die Brauchbarkeit ganzer Therapiesysteme im Vergleich geben und gelten mittlerweile als sehr nützlich.

Forderungen von Ergebnisdarstellungen, die nicht immer mit der Publikationspraxis vereinbar sind

Das Methodenpapier fordert komplette Ergebnisdarstellungen von Statistiken und Outcomes. Das ist im Prinzip verständlich, nachvollziehbar und im Grunde auch gut. Aber es widerspricht der Publikationspraxis. Wenn eine Therapiestudie in einem medizinisch ausgerichteten Journal publiziert wird, z.B. BMJ, Lancet, JAMA Psychiatry, etc., dann werden Gutachter und Editoren dieses Journals den CONSORT-Statements zur Berichtlegung von klinischen Studien folgen und Autoren sind gehalten, eine Liste abzugeben, auf der sie ankreuzen müssen, wo sie welches Item wie berücksichtigt haben. Teil dieses CONSORT-Statements sind auch bestimmte Vorgaben zum Berichten von Statistiken. Dazu gehört z.B., dass bei Baseline-Messungen keine statistischen Tests gemacht werden, dass bei Outcome-Kriterien 95%-Konfidenzintervalle und keine Effektgrößen berichtet werden sollen, und je nach Journal manches andere mehr oder weniger. Ich habe schon erlebt, dass Journal-Gutachter oder Editoren explizit verlangten, die Angabe von Effektgrößen zu streichen.

Wird eine Studie im Rahmen einer Bayesianischen Statistik ausgewertet, dann kommen sowieso keine Signifikanz-Tests vor, weil diese nur Teil der Fisherschen Statistik sind. Im Übrigen ist das Ritual der Signifikanztests schon so oft kritisiert worden, dass es Wunder nimmt, dass dieser Begriff in einem solchen Papier als Kriterium auftaucht (Gigerenzer, 2004; Ioannidis, 2005; Lambdin, 2012; Sedlmeier, 1996).

Die Fülle von Forderungen, die hier als Kriterien aufgestellt werden, ist zwar für eine monographische Darstellung realistisch, in der man sehr detailliert berichten kann. Sie ist unrealistisch für eine gute, peer-reviewte Journal Publikation, bei der in der Regel die Anzahl der Tabellen, die Anzahl der Worte und die Anzahl der Abbildungen begrenzt sind. Selbstverständlich kann ein geschickter Autor auch hier noch mehr Information unterbringen als ein weniger erfahrener, aber die Wunschliste der Information des Methodenpapiers bleibt aus meiner Sicht unerfüllbar, wenn man peer-reviewte Publikationen als Bewertungsbasis zugrunde legt. Sie ist erfüllbar, wenn man längere Berichte annimmt. Diese sind dann aber nicht begutachtet.

Es wäre wünschenswert, dass der Wissenschaftliche Beirat Psychotherapie konkrete, publizierte Beispiele von Studien vorlegt und deren Bewertung, zusammen mit einer ausführlich kommentierten Begründung, die zur Akzeptanz eines Verfahrens geführt haben, um zu demonstrieren, dass die Anwendung dieser Liste tatsächlich realistisch ist.

Fazit

Das Methodenpapier des Wissenschaftlichen Beirats ist in mancher Hinsicht nützlich, aber an entscheidenden Punkten fehlerhaft und inkonsistent. Es macht Voraussetzungen, die nicht reflektiert sind. Einige dieser Voraussetzungen sind sachlich falsch, andere sind nicht belegbar. Dazu gehören die Annahme, dass man spezifische von unspezifischen Effekten trennen kann und muss, sowie dass diese Trennung einem additiven Modell folgen kann. Dies ist nicht belegt und wo untersucht mit großer Wahrscheinlichkeit falsch. Zu diesen unreflektierten Voraussetzungen gehört auch die Annahme, dass man interne Validität über die externe stellen kann, bzw. dass man beide Arten der Validität gleichzeitig maximieren kann. Diese Annahme ist falsch. Die beiden Validitätsformen sind inkompatibel. Keine ist wichtiger als die andere. Daher müssen unterschiedliche Studientypen verwendet und gemeinsam betrachtet werden. Die Überbewertung der Randomisation, zumal in kleinen Studien, trägt nicht zu einer Stärkung der Erkenntnislage bei.

Außerdem enthält der Kriterienkatalog eine Reihe von Widersprüchlichkeiten und Inkonsistenzen: Die Bewertung von interner und externer Validität widersprechen sich. Das Prinzip der reichhaltigen Messung und der Definition eines Hauptzielkriteriums widersprechen sich. Die Forderung einer konventionellen Rekrutierung und die Forschungsethik sind nicht miteinander vereinbar. Die gleichzeitige Forderung von Manualisierung und Therapiepraxis wie im natürlichen Umfeld ist widersprüchlich. Der Kriterienkatalog benachteiligt wichtige Designs. Schließlich ist die Fülle der geforderten Ergebnispräsentation nicht mit der Publikationspraxis und den dort aufgestellten Forderungen kompatibel, zumindest nicht im Bereich der hochrangigen medizinischen Journale.

Zusammengenommen lässt dieses Methodenpapier große Lücken der Methodenreflexion und auch der internen Konsistenz erkennen. Es ist daher aus meiner Sicht nicht geeignet zu einer fairen Bewertung eingereichter Therapiestudien beizutragen und sollte grundlegend überarbeitet werden. Es sollte außerdem eine Beispieldokumentation anhand einiger Studien der Richtlinienverfahren erstellt werden, an der ersichtlich wird, dass Studien der bereits akzeptierten Verfahren diesen Kriterienkatalog erfüllen würden. Erst dann wäre eine Vergleichsgerechtigkeit gegeben.

Der Autor

Harald Walach hat Psychologie in Freiburg studiert und mit dem Diplom 1984 abgeschlossen. 1992 wurde er in Basel in Klinischer Psychologie promoviert. In der Zwischenzeit machte er eine Ausbildung in Psychosynthese, einem nicht anerkannten und nicht zur Diskussion stehenden Verfahren. Dieses hatte er in Einzelberatung und in Gruppenarbeit in der Erwachsenenbildung eingesetzt. Zwischen 1987 und 1994 war er in therapeutischer Supervision bei einem Psychoanalytiker Jungscher und Freudscher Ausrichtung.

Parallel zu seinem Psychologiestudium hat er Philosophie studiert mit dem Schwerpunkt Wissenschaftstheorie und dieses 1994 mit einer zweiten Promotion in Wissenschaftstheorie und Wissenschaftsgeschichte in Wien abgeschlossen.

1998 wurde er in Freiburg habilitiert mit einer Arbeit „Die Bedeutung unspezifischer Therapie-Effekte – Das Beispiel Homöopathie“. Im Rahmen dieser und anschließender Arbeiten hat er sich vor allem mit der Evaluation komplexer Therapieverfahren im Rahmen der Komplementärmedizin, mit der Frage der Bedeutung unspezifischer Therapie-Effekte, die in der Medizin unter dem Namen „Placebo-Effekte“ gehandelt werden und mit entsprechenden Methodenfragen beschäftigt.

Er hat selber klinische Studien durchgeführt, große Kohortenstudien, randomisierte experimentelle Studien, Studien zur Entwicklung und Validierung von Fragebögen, einfache Befragungen, katamnestische Studien, Meta-Analysen, Einzelfallexperimente, Zeitreihenanalysen, Regressionsmodellierungen, eigentlich alle Typen von Studien, die in der klinischen Forschung vorkommen. Er ist mit derzeit 214 peer-reviewten Originalarbeiten in der internationalen Literatur ausgewiesen und belegt damit in der internationalen Zitationsdatenbank von John Ioannidis (Ioannidis, Baas, Klavans, & Boyack, 2019) einen Rang im oberen Drittel der 100.000 weltweit meist zitierten Autoren. Das sind 5 Promille aller wissenschaftlich tätigen Autoren weltweit. Es gibt nur wenige Psychologen in Deutschland, die häufiger zitiert werden. Er hat für alle großen medizinischen Journale und viele klinisch-psychologische Zeitschriften Gutachten erstellt und kennt sowohl die Publikationspraxis als auch die anzuwendenden Kriterien sehr gut, sowohl aus der Perspektive eines Gutachters, als auch aus der eines Autors.

Er ist durch langjährige Mitgliedschaft in der Kommission D des Bundesinstitutes für Arzneimittel mit Zulassungsregularien und der Gesetzeslage des Arzneimittelgesetzes gut vertraut. Seine Mitgliedschaft in der European Association of Psychotherapy ließ ihn am Puls der politischen Geschehnisse um die Zulassung von Psychotherapie bleiben.

Er ist durch seine Themenstellung, der Evaluation komplementärer medizinischer Verfahren, gezwungen gewesen, diese Forschung methodisch auf höchstem Niveau durchzuführen. Außerdem hat ihn diese Arbeit zu einer vertieften Methodenreflexion gezwungen, weil sie dauernd im Spannungsfeld zwischen herrschender medizinischer Methodik, die vor allem der Zulassungslogik neuer Präparate geschuldet ist, und sachgebundener Anforderung fairer Evaluation komplexer Verfahren stand.

Literatur

Aickin, M. (1983). Some large trial properties of minimum likelihood allocation. *Journal of Statistical Planning and Inference, 8*, 11-20.

Aickin, M. (2001). Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. *Journal of Statistical Planning and Inference, 94*, 97-119.

Aickin, M. (2002). Beyond randomization. *Journal of Alternative and Complementary Medicine, 8*, 765-772.

Atmanspacher, H., Römer, H., & Walach, H. (2002). Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics, 32*, 379-406. doi:10.1023/A:1014809312397

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal, 312*, 1215-1218.

Concato, J. (2012). Is it time for medicine-based evidence? *Journal of the American Medical Association, 307*, 1641-1643.

Concato, J., & Horwitz, R. I. (2004). Beyond randomised versus observational studies. *Lancet, 363*, 1660-1661.

Concato, J., Lawler, E. V., Lew, R. A., Gaziano, J. M., Aslan, M., & Huang, G. D. (2010). Observational methods in comparative effectiveness research. *American Journal of Medicine, 123*, e16-e23.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*, 1887-1892.

Conferences on Therapy. (1946). The use of placebos in therapy. *New York Journal of Medicine, 46*, 1718-1727.

Conferences on Therapy. (1954). How to evaluate a new drug. *American Journal of Medicine, 17*, 722-727.

Diener, H. C., Kronfeld, K., Boewing, G., Lungenhausen, M., Maier, C., Molsberger, A., . . . Group, f. t. G. M. S. (2006). Efficacy of acupuncture for the prophylaxis of migraine: A multicentre randomised controlled clinical trial. *Lancet Neurology, 5*, 310-316. doi:DOI:10.1016/S1474-4422(06)70382-9

Djulbegovic, B., Kumar, A., Glasziou, P. P., Perera, R., Reljic, T., Dent, L., . . . Chalmers, I. (2012). New treatments compared to established treatments in randomized trials. *Cochrane Database of Systematic Reviews*(10), Art. No.: MR000024. doi:DOI: 10.1002/14651858.MR000024.pub3.

Frank, J. D. (1961). *Persuasion and Healing: A Comparative Study of Psychotherapy*. Baltimore: Johns Hopkins University Press.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*, 587-606.

Gøtzsche, P. C. (2013). *Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Health Care*. London: Radcliff.

Gøtzsche, P. C. (2015). *Deadly Psychiatry and Organised Denial*. Copenhagen: People's Press.

Grawe, K. (1998). *Psychologische Therapie*. Göttingen: Hogrefe.

Grawe, K. (1999). Gründe und Vorschläge für eine Allgemeine Psychotherapie. *Psychotherapeut, 44*, 350-359.

Grawe, K., Donati, R., & Bernauer, F. (1994). *Psychotherapie im Wandel. Von der Konfession zur Profession*. Göttingen: Hogrefe.

Grawe, K., & Grawe-Gerber, M. (1999). Ressourcenaktivieriung: Ein primäres Wirkprinzip der Psychotherapie. *Psychotherapeut, 44*, 63-73.

Haake, M., Muller, H. H., Schade-Brittinger, C., Basler, H. D., Schafer, H., Maier, C., . . . Molsberger, A. (2007). German Acupuncture Trials (GERAC) for chronic low back pain: randomized, multicenter, blinded, parallel-group trial with 3 groups. *Archives of Internal Medicine, 167*(17), 1892-1898.

Hartmann-Kottek, L. (2021). *Allgemeine Psychotherapie: Schulenübergreifende Wirkprinzipien und gemeinsame Theorieaspekte*. Berlin: Springer.

Hofmann, L., & Walach, H. (2011). Spirituality and religiosity in psychotherapy – A representative survey among German psychotherapists. *Psychotherapy Research, 21*, 179-192. doi:DOI: 10.1080/10503307.2010.536595

Howick, J., Koletsi, D., Ioannidis, J. P. A., Madigan, C., Pandis, N., Loef, M., . . . Schmidt, S. (2022). Most healthcare interventions tested in Cochrane Reviews not effective according to high quality evidence: a systematic review and meta-analysis. *Journal of Clinical Epidemiology, 148*, 160-169. doi:10.1016/j.jclinepi.2022.04.017

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.
Ioannidis, J. P. A., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS Biology, 17*(8), e3000384. doi:10.1371/journal.pbio.3000384
Kaptchuk, T. J. (1998). Intentional ignorance: A history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine, 72*, 389-433.
Kaptchuk, T. J. (2001). The double-blind randomized controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology, 54*, 541-549.
Keats, A., & Beecher, H. K. (1950). Pain relief with hypnotic doses of barbiturate and a hypothesis. *Journal of Pharmacology and Experimental Therapeutics, 100*, 1-13.
Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the food and drug administration. *PLoS Medicine, 5*(2), e45. doi:DOI: 10.1371/journal.pmed.0050045
Kirsch, I., Huedo-Medina, T. B., Pigott, H. E., & Johnson, B. T. (2018). Do outcomes of clinical trials resemble those "real world" patients? A reanalysis of the STAR*D antidepressant data set. *Psychology of Consciousness: Theory, Research, and Practice, 5*(4), 339-345. doi:10.1037/cns0000164
Klement, R. J., Bandyopadhyay, P. S., Champ, C. E., & Walach, H. (2018). Application of Bayesian evidence synthesis to modelling the effect of ketogenic therapy on survival of high grade glioma patients. *Theoretical Biology and Medical Modelling, 15*(12). doi:https://doi.org/10.1186/s12976-018-0084-y
Klibansky, R. (1936). Standing on the shoulders of the giants. *Isis, 26*, 147-149.
Koch, A., Hörmann, A., Löwel, H., & Senges, J. (1998). Problems of randomized trials. In U. Abel & A. Koch (Eds.), *"The 60minutes-myocardial infarction project": Comparison with a registry and a randomized trial* (pp. 149-160). Düsseldorf: Symposion Publishing.
Kriz, J. (2000). Perspektiven der "Wissenschaftlichkeit" in der Psychotherapie. In M. Hermer (Ed.), *Psychotherapeutische Perspektiven am Beginn des 21.Jahrhunderts* (pp. 43-66). Tübingene: DGVT-Verlag.
Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology, 22*(1), 67-90. doi:10.1177/0959354311429854
Lilford, R. J., & Jackson, J. (1995). Equipoise and the ethics of randomization. *Journal of the Royal Society of Medicine, 88*, 552-559.
Linde, K., Scholz, M., Melchart, D., & Willich, S. N. (2002). Should systematic reviews include non-randomized and uncontrolle studies? The case of acupuncture for chroic headache. *Journal of Clinical Epidemiology, 55*, 77-85.
London, A. (2017). Equipoise in research: Integrating ethics and science in human research. *JAMA, 317*(5), 525-526. doi:10.1001/jama.2017.0016
Martini, P. (1932). *Methodenlehre der therapeutischen Untersuchung.* Berlin: Springer.
McQuay, H., Carroll, D., & Moore, A. (1996). Variation in the placebo effect in randomised controlled trials of analgesics: all is blind as it seems. *Pain, 64*, 331-335.
Meissner, K., Fässler, M., Kleijnen, J., Hróbjartsson, A., Schneider, A., Antes, G., & Linde, K. (2013). Differential effectiveness of placebo treatments: A systematic review of migraine prophylaxis. *JAMA Internal Medicine, 173*, 1941-1951.
Miller, F. G., & Joffe, S. (2011). Equipoise and the dilemma of randomized clinical trials. *New England Journal of Medicine, 364*, 476-480.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V. M., Gøtzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c869. doi:10.1136/bmj.c869

Pigott, H. E., Leventhal, A. M., Alter, G. S., & Boren, J. J. (2010). Efficacy and effectiveness of antidepressants: current status of research. *Psychotherapy and Psychosomatics, 79*, 267-279.

Rief, W., & Gaab, J. (2016). Die dunkle Seite der Intervention - was hat Placebo mit Psychotherapie zu tun? *Verhaltenstherapie, 26*, 6-7.

Rush, J. A., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., . . . Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report *American Journal of Psychiatry, 163*, 1905-1917

Scharf, H.-P., Mansmann, U., Streitberger, K., Witte, S., Krämer, J., Maier, C., . . . Victor, N. (2006). Acupuncture and knee osteoarthritis. *Annals of Internal Medicine, 145*, 12-20.

Schulz, K. F., Altman, D. G., Moher, D., & CONSORT Group. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine, 8*(1), 18. doi:10.1186/1741-7015-8-18

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Reserach Online, 1*(4).

Seligman, M. (1995). The effectiveness of psychotherapy. *American Psychologist, 50*, 965-974.

Trivedi, M. H., Fava, M., Wisniewski, S. R., Thase, M. E., Quitkin, F., Warden, D., . . . Team, S. D. S. (2006). Medication augmentation after the failure of SSRIs for depression. *New England Journal of Medicine, 354*(12), 1243-1252.

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., . . . Team, S. D. S. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *American Journal of Psychiatry, 163*(1), 28-40.

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.

Walach, H. (2001). Das Wirksamkeitsparadox in der Komplementärmedizin. *Forschende Komplementärmedizin und Klassische Naturheilkunde, 8*, 193-195.

Walach, H. (2011). Placebo controls: historical, methodological and general aspects. *Philosophical Transactions of the Royal Society Biological Sciences, 366*, 1870-1878.

Walach, H. (2016). The efficacy paradox and its consequences for research in psychotherapy (and elsewhere). *Psychology of Consciousness: Theory, Research, and Practice, 3*(2), 154-161.

Walach, H. (2020, orig. 2005). *Psychologie: Wissenschaftstheorie, philosophische Grundlagen und Geschichte* (5. überarb. Aufl. ed.). Stuttgart: Kohlhammer.

Walach, H., Falkenberg, T., Fonnebo, V., Lewith, G., & Jonas, W. (2006). Circular instead of hierarchical - Methodological principles for the evaluation of complex interventions. *BMC Medical Research Methodology, 6*(29). doi:doi.org/10.1186/1471-2288-6-29

Walach, H., & Loef, M. (2015). Using a matrix-analytical approach to synthesizing evidence solved incompatibility problem in the hierarchy of evidence. *Journal of Clinical Epidemiology, 68*, 1251-1260. doi:doi:10.1016/j.jclinepi.2015.03.027

Walach, H., & Loughlin, M. (2018). Patients and agents – Or why we need a different narrative: A philosophical analysis. *Philosophy, Ethics and Humanities in Medicine, 13*(13). doi:10.1186/s13010-018-0068-x

Walach, H., Sadaghiani, C., Dehm, C., & Bierman, D. J. (2005). The therapeutic effect of clinical trials: understanding placebo response rates in clinical trials - A secondary analysis. *BMC Medical Research Methodology, 5*, 26.

Wampold, B. E. (2021). *Evaluation: Methodology Paper of the Scienctific Advisory Board on Psychotherapy According to Section 11 PsychThG (Psychotherapists Act)*. Retrieved from Kassel: https://ddgap.de/cms/wp-content/uploads/2021/06/Evaluation-Methods-Paper-Final-2.pdf

Wampold, B. E., Frost, N. D., & Yulish, N. E. (2016). Placebo effects in psychotherapy: A flawed concept and a contorted history. *Psychology of Consciousness: Theory, Research, and Practice, 3*(2), 108-120.

Wampold, B. E., & Imel, Z. E. (2015). *The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work*. London: Routledge.

West, C. P., Dyrbye, L. N., Rabatin, J. T., Call, T. J., Davidson, J. H., Multari, A., . . . Shanafelt, T. D. (2014). Intervention to promote physician well-being, job satisfaction and professionalism: A randomized clinical trial. *JAMA Internal Medicine, 174*(4), 527-533. doi:10.1001/jamainternmed.2013.14387

Whitehead, A. N. (1978). *Process and Reality. Corrected Edition by D.R. Griffin & D.W. Sherburne. First Ed. 1929*. New York: Free Press.

Wittmann, W. W. (1985). *Evaluationsforschung. Aufgaben, Probleme und Anwendungen*. Berlin: Springer.

Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of Multivariate Experimental Psychology* (2nd edition ed., pp. 505-560). New York: Plenum Press.

Zwarenstein, M., Treweek, S., Gagnier, J. J., Altman, D. G., Tunis, S., Haynes, B., . . . Group, C. P. T. i. H. (2008). Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *British Medical Journal, 337*, a2390. doi:doi: 10.1136/bmj.a2390