Einige einführende Gedanken zum Verständnis unserer neuen Publikation

> *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:10.1016/j.jclinepi.2015.03.027*

Bevor ich ein paar kritische Bemerkungen zur „Evidence Based Medicine (EBM)“ mache, ist es mir ein Anliegen, nochmals diese Bewegung zu loben, für all das, was sie erreicht hat, was ihre Ziele sind und was ihre Vision ist: eine medizinische Versorgung gestützt auf wissenschaftliche Befunde, nicht auf Autoritätsmeinungen; das Ablösen von Eminenz durch „Evidenz“ (auch wenn „evidence“ im Englischen etwas anderes bedeutet als im Deutschen, nämlich „Beweis“; im Deutschen hingegen ist „evident“ genau das, was keines Beweises bedarf). Das ist ein aufklärerischer, befreiender Impuls, der sehr gut und wichtig ist.

Viele, die sich dieser Bewegung angeschlossen haben, tun dies jedoch unreflektiert. Sie übersehen die subtile Dialektik, die sich in fast allen historischen Prozessen äußert. Das, was man meint überwunden zu haben – hier: Autoritätshörigkeit – kehrt in einem etwas neuem Gewand wieder zurück. Denn das, was früher die persönliche Autorität des Chefarztes war, der sagte wo’s lang geht, ist nun die anonyme Autorität der Leitlinien. Und diese speisen sich vor allem aus bereits publizierten systematischen Übersichtsarbeiten und Meta-Analysen. Und in diese gehen wiederum fast ausschließlich randomisierte Studien ein. Der Rest aller Daten wird einfach ignoriert. Das ist verständlich. Denn wir Menschen sind faule Wesen und gehen gern den kürzesten Weg und den, bei dem es am wenigsten Widerstand gibt.

Dahinter steht natürlich eine, wie es zunächst aussieht, kluge Idee: Studiendesigns kann man in eine hierarchische Ordnung bringen, die die Verlässlichkeit der Schlussfolgerungen spiegelt, die man aus ihnen ziehen kann. Daraus ergibt sich die berühmte „Evidenz-Hierarchie“ der EBM. Unten stehen die einfachen Beobachtungen: Fallserien, gut dokumentierte Kohorten, aber ohne direkten Vergleich. Den muss man sich aus seinem Wissen oder aus bekannten Daten selber generieren. Eine Stufe höher stehen Studiendesigns mit parallelen Vergleichsgruppen, sogenannte Kohortenstudien, oder naturalistische Vergleiche, die auf natürliche Weise zustande gekommen sind. Dies sind etwa Gruppen von Patienten, die sich ihre Behandlung selber ausgesucht haben und dann

miteinander verglichen werden. Wiederum eine Stufe höher stehen randomisierte Studien. Das sind solche Studien, bei denen der Experimentator die Gruppen erzeugt hat, in dem er Patienten durch einen Zufallscode auf die Gruppen aufteilte. Wenn man mehrere solcher Studien hat, kann man sie in Meta-Analysen oder systematischen Überblicksarbeiten zusammenfassen und hat das „wahre“ Wissen. So in etwa die Theorie.

Diese ist mittlerweile etwas differenzierter [1]. Man kann nämlich naturalistische Studien aufwerten, wenn sie besonders gut oder besonders groß sind und randomisierte Studien abwerten, wenn sie sehr klein oder organisatorisch schlecht gemacht sind. Aber das Prinzip und die Denkhaltung sind die gleichen:

Weil diese Hierarchie eine Hierarchie der Zuverlässigkeit von Schlußfolgerungen ist, die man aus den Studien ziehen kann, technisch gesprochen: *eine Hierarchie der internen Validität, darum kann man Informationen aus einer tieferen Hierarchie-Ebene ignorieren, sobald man bessere Informationen hat*. Daher werden in Meta-Analysen und systematischen Reviews fast ausschließlich randomisierte Studien berücksichtigt. *Das habe ich schon immer für einen kapitalen Fehler gehalten,* und das haben wir in diesem Aufsatz argumentativ belegt und außerdem gezeigt, dass und wie es anders geht.

Bevor wir auf ein paar Gründe eingehen, warum ich dieses Mainstream-Narrativ für kurzsichtig halte, ist eine Erinnerung wichtig: einer der Gründerväter der Evidence Based Medicine, David Sackett [2], hat immer betont, dass es *drei* Pfeiler sind, auf denen die EBM ruht. 1. die bestmöglichen, wissenschaftlichen Daten, 2. die klinische Erfahrung des Arztes und 3. die Präferenz des Patienten. Sehr oft wird nur noch die wissenschaftliche Datenbasis berücksichtigt und der Rest ignoriert.

Nun ist es wichtig zu verstehen, dass die *Gültigkeit der Schlußfolgerungen, die interne Validität* einer Studie, *nur eine Form* der Validität ist. Es gibt eine *andere Validitätsart*, die aus meiner Sicht genauso wichtig ist, die *externe Validität*, oft auch als Generalisierbarkeit bezeichnet. Streng genommen gibt es noch zwei weitere, die ökologische und die Modell-Validität, aber die ignorieren wir jetzt. Die externe Validität gibt an, ob und für wen die Ergebnisse einer Studie generalisierbar sind. Das Mainstream-Narrativ geht davon aus, dass interne und externe Validität gleichsam linear additiv sind und sich gegenseitig implizieren: so, als müsste erst interne Validität gegeben sein, damit man sich um die externe kümmern kann (oder auch nicht). Wir zeigen in dem Aufsatz, dass diese Haltung faktisch und logisch falsch ist. Alle Gründe kann ich hier nicht aufführen, ohne das ganze Paper auf Deutsch nochmals zu schreiben. Aber ein paar seien genannt:

1. Um eine experimentelle klinische Studie, also eine randomisierte intern hoch valide Studie zu erzeugen, muss man experimentelle Kontrolle behalten. Man versucht Studiengruppen zu homogenisieren – indem man Einschluss- und Ausschlusskriterien

definiert und anwendet. Dies erhöht idealer Weise den Unterschied zwischen den Gruppen und macht die Trennschärfe grösser. Technisch gesprochen: man erhöht systematische Varianz und reduziert Fehlervarianz. Das ist sozusagen implizit im experimentellen Vorgehen. Je mehr solcher Ein- und Ausschlusskriterien man verwendet, desto „sauberer“ wird die Studie. Ihre interne Validität steigt. Aber gleichzeitig sinken Generalsierbarkeit und externe Validität.

2. Um eine experimentelle klinische Studie durchführen zu können, müssen Patienten einwilligen. Viele tun dies nicht. Diejenigen, die einwilligen, das wissen wir aus Studien, sind nicht mit denen vergleichbar, die nicht einwilligen. Also gelten die Ergebnisse nicht für die Patienten, die nicht einwilligen und solche, die diesen ähnlich sind.
3. Viele randomisierte Studien, vor allem solche zu Zulassungszwecken, schränken die Behandlungszeit ein, weil längere Studien teurer sind, und sie schränken auch die Patientengruppen ein. Mit beiden Maßnahmen erhöhen sie die Chance, die Wirksamkeit einer Intervention zu belegen und steigern die interne Validität. Damit verlieren sie an externer Validität.
4. Randomisierte Studien müssen per definitionem ihre Studienteilnehmer als passive Empfänger therapeutischer Leistungen konzipieren. Das mag für Pharmazeutika angehen. Das ist aber falsch für alle komplexen Interventionen, bei denen die Aktivität und Mitarbeit von Patienten gefragt ist. Daher können randomisierte Studien immer nur die minimal zu erwartende Effektgröße schätzen und versagen, wenn es darum geht, maximale Effekte zu erzeugen. Denn dafür benötigt man Studiendesigns, die den Patienten Wahlfreiheit lassen und die auch den Behandlern die Option offen lassen diejenige Intervention zu wählen, von der sie am meisten überzeugt sind.

All dies sind Domänen der „externen Validität“ oder der Generalisierbarkeit von Studienergebnissen. Diese ist höher in naturalistischen Studien, bei denen Patienten und Behandler aussuchen können, was sie tun möchten. In gewisser Weise verbergen sich dahinter die beiden anderen, vernachlässigten Pfeiler der EBM im Sinne Sacketts, die Wahlmöglichkeit des Patienten und die klinische Erfahrung des Arztes.

Der Punkt unseres Argumentes ist nun: *interne und externe Validität sind nicht miteinander kompatibel*; sie schließen sich in gewisser Weise aus. Jede Studie, die die interne Validität erhöht, verringert die externe Validität, und umgekehrt. *Es ist keine Studie denkbar, und zwar aus prinzipiellen Gründen, die beides, externe und interne Validität gemeinsam erhöht*, und ich habe auch faktisch noch keine gesehen, bei der das der Fall wäre. Auch die Reviewer des Aufsatzes, denen ich die Herausforderung herübergereicht habe, mir eine zu zeigen, mussten passen.

Daher sind interne und externe Validität inkompatible Konzepte. Und wer sich ein bisschen mit dem Formalismus unserer generalisierten Quantentheorie beschäftigt hat, sieht sofort, dass man zur theoretischen Modellierung solcher Konzepte einen anderen Formalismus als den klassischen, linear-additiven benötigt [3]. Daraus folgt im übrigen: die Reihenfolge, in der man Erkenntnis erzeugt, ist nicht egal.

Praktisch sieht man das an zwei Beispielen. In der Erfahrungsheilkunde gibt es oft sehr viel Erfahrungswissen – anwendbar, generalisierbar – das aber keine sonderlich hohe methodische Gültigkeit hat und nicht auf intern valide Weise gewonnen wurde. Aber das Wissen ist vorhanden. Und wenn dann harte Studien Ergebnisse produzieren, die mit diesem Wissen nicht kompatibel sind, dann wird diese neue Erkenntnis meistens ignoriert.

Daher ist die Forschungslogik in der Komplementärmedizin genau umgekehrt wie in der konventionellen, auf pharmakologische Substanzen basierten Forschung [4]. Umgekehrt erzeugen intern hoch valide Studien oft gute Daten zur Wirksamkeit neuer Interventionen. Aber Wissen um deren Generalisierbarkeit und Anwendbarkeit kommt oft viel später. Darum kann es sein, dass erst nach Jahren der Zulassung gefährliche Nebenwirkungen, Interaktionen, oder das Wissen um eine sehr begrenzte Nützlichkeit erzeugt wird. Meistens kommt dies dann aber zu spät. Denn die Zulassung ist ausgesprochen, das Produkt in Anwendung und es muss viel passieren, bis Indikationseinschränkungen oder Rücknahmen durch die Behörden erfolgen. Man sieht daran: die *Reihenfolge*, in der das Wissen erzeugt wurde, spielt eine Rolle. Technisch gesprochen: *interne und externe Validität kommutieren nicht*.

Was heißt das konkret?

Daraus lassen sich aus meiner Sicht folgende Konsequenzen ableiten:

1. Es ist nicht ausreichend, nur die vermeintlich „besten“ Daten, also randomisierte Studien heranzuziehen. Ihre Aussagen müssen mit den Aussagen von naturalistischen Studien kontrastiert werden.
2. Forschungsmethoden sind nicht hierarchisch aufeinander bezogen, sondern eher zirkulär oder wie in einem Mosaik: sie ergänzen sich gegenseitig [5].
3. Daher sollten in systematischen Reviews Daten aller Studientypen berücksichtigt werden und in ihren Aussagen gegeneinander gestellt werden. Wir haben das als matrix-analytischen Ansatz bezeichnet. Dabei wird eine Matrix erstellt, die verschiedene Studientypen in den Zeilen abbildet und in den Spalten die Anzahl der Studien oder Effektgrössen, die eine Hypothese stützen oder widerlegen oder unentschieden sind. Wenn man das Gesamtbild überblickt, sieht man rasch, ob es konsistente Trends gibt oder Widersprüche. Solche müsste man auflösen, entweder indem man die Studien genauer unter die Lupe nimmt, oder sich modifizierende

Hypothesen überlegt.

4. Auf jeden Fall sollte die „quick-and-dirty“-Methode systematischer Reviews aufhören [6]: ich such‘ mir die randomisierten Studien und ignoriere den Rest und meine damit einen fairen, vollständigen und wissenschaftlich akzeptablen Überblick über die Literatur gegeben zu haben. Was dabei allenfalls herauskommt sind Publikationen, die man sich an den Hut stecken kann. Forschungsförderer, Gutachter, Reviewer, die Cochrane Collaboration – alle sollten damit aufhören, solche Reviews zu fördern, zu schreiben und zu publizieren.

Quellen

1. Howick, J. (2011). *The Philosophy of Evidence-Based Medicine*. Chichester: Wiley-Blackwell.
2. Sackett, D. L. (1997). *Evidence Based Medicine: How to Practice and Teach EBM*. New York: Churchill Livingstone.
3. Filk, T., & Römer, H. (2011). Generalized Quantum Theory: Overview and latest developments. *Axiomathes, 21*, 211-220.
   Walach, H., & Stillfried, N. v. (2011). Generalised Quantum Theory—Basic idea and general intuition: A background story and overview. *Axiomathes, 21*, 185-209.
4. Fonnebo, V., Grimsgaard, S., Walach, H., Ritenbaugh, C., Norheim, A. J., MacPherson, H., et al. (2007). Researching complementary and alternative treatments – the gatekeepers are not at home. *BMC Medical Research Methodology, 7*(7). www.biomedcentral.com/1471-2288/7/7
5. 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). http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-6-29 Siehe auch: http://www.altmetric.com/details/733156
6. Vickers, A. J. (2010). Reducing systematic reviews to a cut and paste. *Forschende Komplementärmedizin, 17*, 303-305.