Nur etwa 6 Prozent aller in der Medizin angewandten Interventionen haben eine ausreichend gute Datenlage und sind wirksam

Unser neuer Meta-Review zeigt: Das Rückgrat der Evidence Based Medicine ist schwach

Regelmäßige Leser meiner Texte wissen, dass ich sehr skeptisch gegenüber dem postmodernen Erlösungsnarrativ der modernen Medizin bin, das da verkündet: *Wir leben so lange und es geht uns so gut, weil die moderne Medizin so gewaltige Fortschritte gemacht hat. Daher ist alles, was uns die moderne Pharmakologie beschert, gut, begrüßens- und unterstützenswert (und sollte durch die Öffentlichkeit finanziert werden)*.

Schon der legendäre Sozialmediziner Thomas McKeown aus Birmingham hat in den 70er Jahren darauf hingewiesen, dass diese verbreitete landläufige Meinung höchstwahrscheinlich falsch ist und meinte in der Einleitung zu seinem immer noch sehr lesenswerten Werk „The Role of Medicine: Dream, Mirage, or Nemesis? – Die Bedeutung der Medizin: Traum, Trugbild, oder Nemesis?“ [1,2]: Wenn er Petrus wäre, er würde nur zwei Typen von Ärzten in den Himmel lassen, nämlich Unfallchirurgen und Zahnärzte. Denn das wären die einzigen, die wirklich zu einer Verringerung von Leiden beigetragen hätten. Die eigentlichen Fortschritte und damit die Verlängerung der Lebensspanne und die Verbesserung der Lebensqualität würden wir nicht der Medizin verdanken, sondern sozial-politischen Fortschritten, besserer Ernährung, der Hygiene und Lebensbedingungen ohne dauernde Angst vor Not und Tod.

Nun, das war in den 70ern. Vielleicht ist es heute anders? Wir haben eine sehr groß angelegte Meta-Studie durchgeführt, um die Frage zu beantworten, wie gut die Datenlage für medizinische Interventionen im Allgemeinen ist. Sie ist jetzt im Journal of Clinical Epidemiology publiziert worden [3]. Ich diskutiere in diesem Blog die Studie und ihre Befunde etwas genauer. Für Eilige: Die Datenlage hat sich nicht groß geändert. Maximal 6 Prozent aller in der Medizin angewandten Intervention, egal wo, sind durch gute Datenlage gedeckt.

Nur kurze Zeit (bis 22.08.2022) ist der Meta-Review frei zugänglich unter: https://authors.elsevier.com/c/1fIHz3BcJQAobl

Die Studie wurde initiiert von Jeremy Howick, der lange am Oxford Center of Evidence

Based Medicine gearbeitet hat und sich sehr intensiv um die konzeptuelle Durchdringung der Evidence Based Medicine (EBM) gekümmert hat [4]. Evidence Based Medicine (EBM) müsste man eigentlich korrekterweise mit „Faktengestützte Medizin“ und nicht mit „evidenzbasierte Medizin“ übersetzen. Denn „evidence“ heißt im Englischen „Beweis“, also das, was durch Fakten und Daten untermauert wird, während „evident“ im Deutschen gerade das ist, was keines Beweises bedarf, eben weil es klar ist. Aber das nur am Rande.

In diesem Meta-Review ging es uns darum herauszufinden, wie viele Cochrane Reviews wirklich solide belegte Hinweise auf Wirksamkeit liefern, und zwar quer durch alle Bereiche der Medizin, von Chirurgie bis zu Psychiatrie, von Kinderheilkunde bis zu Gerontologie und Verhaltensinterventionen.

Die Cochrane Collaboration und Cochrane Reviews

Die Cochrane Collaboration ist ein ursprünglich selbst-organisiertes Netzwerk von Forschern, mittlerweile eine Stiftung. Forscher, die dort mitarbeiten, tun das aus wissenschaftlichem Interesse und allenfalls eine kleine Gilde von Hauptamtlichen erhält Finanzierung von lokalen Geldgebern. Das große Markenzeichen der Cochrane Collaboration war Unabhängigkeit von Interessensgruppen und Industriegeldern. Die Cochrane Collaboration ist in sog. Review-Groups aufgeteilt, also Autorengruppen, die sich um bestimmte größere Gebiete, z.B. Kardiologie, Onkologie etc. kümmern und innerhalb dieser wiederum um spezielle Fragen.

Die Cochrane Collaboration publiziert die sog. Cochrane Library, also eine Sammlung aller Reviews, die von den Forschern des Netzwerkes durchgeführt werden. Auf der Webseite sind auch Protokolle verzeichnet, also Definitionen von Reviews und ihrer Methodik, die gerade durchgeführt werden. Und wichtige Reviews werden regelmäßig aktualisiert, wenn neue Daten verfügbar werden.

Damit stellt die Cochrane Library das Herz der EBM dar. Denn sie fasst das an Wissen zusammen, was für einzelne Fachbereiche, Diagnose- oder Behandlungsfragestellungen in der Medizin wirklich wichtig ist. Die Zusammenfassung dieses Wissens geschieht, indem zu einer interessierenden Fragestellung und Intervention alle Studien – vor allem randomisierte klinische Studien, oft auch nicht-randomisierte Kohortenstudien oder Fallkonstrollstudien, je nach Fragestellung, zusammengefasst und am Ende bewertet werden. Wer also z.B. wissen will, ob Ritalin bei Aufmerksamkeits-Hyperaktivitätssyndrom (ADHS) bei Kindern wirkt und empfohlen wird, könnte die Cochrane Library durchsuchen und fände sowohl einen Review über die Wirksamkeit, als auch einen über die Nebenwirkungen von Ritalin [5,6]. Im einen Fall [6] wurden insgesamt 38 Studien mit mehr als 5.000 Kindern eingeschlossen. Die Autoren kamen zur Ansicht, dass alle Studien

Designfehler aufweisen, die sie anfällig für Verzerrungen machen. Daher sei unklar, ob der kleine Effekt, den sie gefunden haben, auch wirklich klinisch bedeutsam ist. Im anderen Fall [5] wurden 260 Studien eingeschlossen, die ein hohes Nebenwirkungspotenzial belegen, das aber schwer quantifizierbar ist.

Diese beiden Beispiele sind ziemlich typisch für Cochrane Reviews; ich erwähne sie auch deshalb, weil ich in meinem letzten Blog auf unsere eigene Meta-Analyse zu Homöopathie bei ADHS eingegangen bin. Die Beispiele zeigen, wie die Autoren vorgehen und wie viel Synthesearbeit dahintersteckt.

Cochrane Reviews sind sehr rigide. Sie folgen einem vor-definierten methodischen Raster und versuchen, möglichst alle Studien zu finden und einzuschließen. Um die Bewertung einfacher zu machen, wurde etwa 2008 das sog. GRADE-System eingeführt: Grading of Recommendations, Assessments, Development and Evaluation, also eine Art Bewertungsschablone [7-9]. Wir haben nun in unserer eigenen Studie ein Drittel aller Cochrane-Reviews untersucht, die dieses GRADE-System angewandt haben.

Die Auswahl der Studien

Da wir uns nur für die Interventionen interessierten, bei denen das GRADE-System zur Bewertung der Datenlage angewandt worden ist, also seit 2008, als es allgemein üblich wurde, war der Gesamtdatensatz der Cochrane-Library seit damals 6.928 Reviews groß. Wenn man alle Interventionen berücksichtigen würde, wären es noch mehr. Das ist natürlich eine riesige Menge von Reviews, die man auch mit einer großen Forschergruppe nicht mehr handhaben kann. Daher entschieden wir uns dafür, ein Drittel davon zufallsgesteuert auszuwählen und diese zu bewerten; auch das noch eine ziemliche Menge Material. Wir waren 12 Forscher und auf jeden entfielen damit etwa knapp 80 Reviews zur Extraktion und Sichtung, im Durchschnitt. Denn von den ausgewählten und infrage kommenden 2.428 Reviews erfüllten nur 1.076 Reviews die Einschlusskriterien. Diese Reviews deckten insgesamt 1.567 Interventionen ab (weil manche Reviews mehrere Interventionen untersuchen). Die ausgeschlossenen Reviews wurden meist deshalb ausgeschlossen, weil sie keine GRADE-Bewertung enthielten, oder weil sie Interventionen mit aktiven Kontrollen verglichen. Wir wollten nämlich nur solche haben, die eine Intervention mit Placebo, keiner Behandlung oder Standard-Therapie verglichen.

Das Vorgehen

Wir extrahierten jeder die ihm zugewiesene Anzahl von Studien, in ein Excel-Spreadsheet, das vorher getestet worden war. Uns interessierte vor allem: Gibt es in dem Review ein Outcome, das den Effekt der Intervention erfasst, das als „high quality“ nach dem GRADE

System bewertet wurde und, in zweiter Linie, gibt es Hinweise auf Nebenwirkungen und Schäden und wenn ja, gibt es dafür einen GRADE Indikator. Wir extrahierten also „high quality" outcomes und ihre Effektgröße, aber auch Nebenwirkungen, und wenn solche dokumentiert waren auch den „GRADE" Indikator, neben einigen anderen interessierenden Größen, z.B. ob gegen Placebo kontrolliert wurde, gegen Standardbehandlung oder keine Behandlung, wie viele Studien in einem Review waren, welche Intervention, welche Population und welche Diagnose untersucht worden war.

GRADE

GRADE ist ein Verfahren, das bei systematischen Reviews die dort zusammengefassten Outcomes einzeln bewertet. Wenn beispielsweise in Lipidsenker-Studien die Blutfettwerte als Zielkriterium erfasst werden und nur in manchen Mortalität (weil sich natürlich Blutfettwerte viel leichter erfassen lassen und rascher verfügbar sind als Mortalitätsdaten), dann würde GRADE das Outcome „Blutfettwerte" als „low quality" oder „very low quality" evidence bewerten. Denn dies sind Surrogatparameter, und wenn die Studie kurz war oder die dokumentierten Effekte klein, dann würde diese Ergebnislage die Bewertung des Outcomes „Blutfettwerte" insgesamt beeinflussen und die Autoren eines solchen Reviews würden vielleicht schreiben „low-quality evidence for effectiveness" oder etwas Ähnliches. Umgekehrt würde Mortalität als Outcome oder Ergebnisparameter, wenn er über einen langen Zeitraum bei ausreichend vielen Patienten einen klinisch bedeutsamen Effekt erkennen lässt als „high quality evidence" bewertet werden. Denn das GRADE-System bewertet nicht nur die Signifikanz, sondern auch die klinische Bedeutsamkeit, die numerische Größe des Effekts und die Angemessenheit des Forschungsdesigns für die Fragestellung, sowie die Frage, ob die untersuchten Patienten für die Fragestellung repräsentativ waren und ob die Effektgrößen weit streuen. (Näheres dazu im Methodenblog zur Meta-Analyse.)

Das GRADE-System dient also dazu, sowohl die wissenschaftliche, als auch die klinisch-praktische Bedeutsamkeit eines Befundes einzuordnen. Uns interessierte: Bei wie vielen Reviews würden wir klare Hinweise darauf finden, dass die einzelnen Studien, die dem Review zugrunde liegen, gute Datenqualität haben, sodass der Review von „high quality of evidence", also von einem klaren Hinweis auf klinische *und* wissenschaftlich belegte Wirksamkeit reden würde.

Das Ergebnis

Unsere Zufallsauswahl brachte uns Reviews von allen 53 Cochrane-Gruppen, also über alle klinischen Fragestellungen hinweg. Die meisten Interventionen wurden in randomisierten Studien getestet. Mehr als die Hälfte aller Interventionen waren pharmakologischer Natur,

16 % waren psychologische oder Verhaltensinterventionen, 6,4 % waren chirurgische Interventionen, und andere Interventionen wie Diät und Ernährung, Bewegung, alternative Therapien machten jeweils nicht mehr als 3 % der Interventionen aus. 45 % aller Interventionen verglichen mit Placebo, 35 % mit Standardtherapie, und der Rest mit Nichtbehandlung.

Das Ergebnis wird am besten mit unserem Schaubild illustriert:

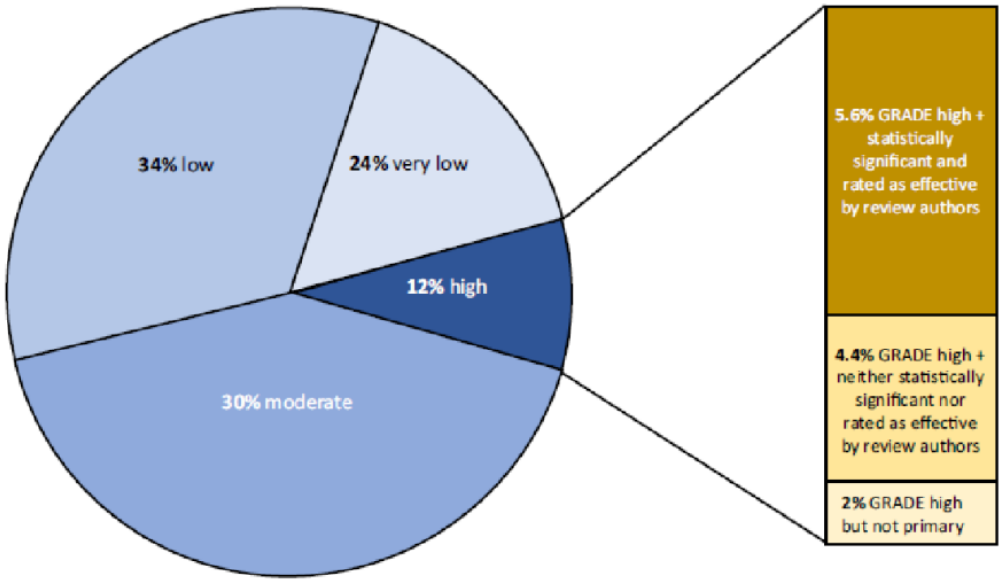

Fig. 1. Proportion of interventions according to their highest GRADE outcome (high, moderate, low, very low).

Abbildung aus der Originalpublikation [3]: Anteil der Interventionen mit hohem, moderaten, niedrigen oder sehr niedrigem GRADE -Rating

5,6 % aller Interventionen hatten sowohl ein Outcome mit einem „high quality“ GRADE Rating *und* einem statistisch signifikanten Effekt (der dunkelbraune Balken). 4,4 % der Reviews berichteten von einem zwar als „high quality“ eingestuften GRADE rating, das aber weder signifikant noch von den Autoren als effektiv bewertet wurde. 2 % hatten zwar ein GRADE Rating, das hoch war, aber lediglich bei einem Nebenzielkriterium. Ein Beispiel dafür wäre etwa, wenn eine onkologische Studie als Hauptzielkriterium das rückfallfreie Zeitintervall nehmen würde, das aber ein Surrogatparameter ist und daher kein „high quality“ Rating erhalten würde und als Nebenzielkriterium die Lebensqualität, das zwar „high quality“, weil klinisch relevant wäre, aber nicht Hauptzielparameter der Studie war.

Man sieht an der Kuchengrafik: Bei 58 % der Reviews ist die Qualität des Hauptzielkriteriums und damit des klinischen Ergebnisses als „low“ oder „very low“ bewertet, bei 30 % als mäßig.

Man kann also festhalten:

Bei weniger als 6 % aller medizinischen Interventionen haben wir einen hohen Grad an klinischer und wissenschaftlicher Sicherheit, dass die Intervention wirksam und klinisch brauchbar ist.

Dies ist, wohlgemerkt, etwas anderes als wissenschaftlich belegte Signifikanz. Denn eine Studie kann Signifikanz erzeugen, die dennoch für die klinische Anwendung irrelevant ist, z.B. weil der Effekt zu klein ist, weil das Zielkriterium klinisch nicht bedeutsam war, weil die Studienpopulation zu speziell war, sodass das Ergebnis nicht übertragbar ist und aus einer Vielzahl von anderen Gründen.

Nebenwirkungen

Diesem bescheidenen Wirksamkeitsprofil stehen die Nebenwirkungen gegenüber. Nur bei 577 oder 37 % aller Interventionen wurden auch die Nebenwirkungen so dokumentiert, dass sie in den Reviews erfasst werden konnten. Nur wenige, nämlich 6 % dieser Studien, hatten für die Nebenwirkungen ein als „high quality“ bezeichnetes Outcome für Nebenwirkungen, z.B. Mortalität. 22 % all der Reviews, die Nebenwirkungen dokumentierten, stellten signifikante Schäden fest, also nicht nur irgendwelche Nebenwirkungen, sondern „harm“, also Schäden, die durch die Intervention verursacht wurden.

Einflussfaktoren

Wir untersuchten mögliche Studienmerkmale, die das Ergebnis beeinflussen könnten – Diagnosegruppen, Studiendesigns, Interventionstypen, etc. – fanden aber keine.

Einschränkungen

Auch wenn wir uns Mühe gaben: Keine Studie ist perfekt. Man könnte zum Beispiel argumentieren, dass es Interventionen gibt, die klarerweise wirksam sind, aber nicht in unser Raster fielen, weil sie schon viel älter sind und nicht durch Studien, die nach 2008 gemacht wurden, untersucht werden müssen, Schienen von Beinbrüchen, Notfall-OPs bei Schlagaderverletzungen oder schweren Unfällen, Antibiotika-Therapie bei bakterieller Pneumonie, Insulinersatztherapie bei Diabetes, Magen auspumpen bei Medikamentenmissbrauch, etc. Das stimmt sicherlich und insofern ist unsere Zahl vielleicht eine – leichte – Unterschätzung. Denn frühere Reviews, die ebenfalls eine große Zufallsauswahl von Cochrane-Reviews untersuchten, ohne die GRADE-Bewertung, die es damals noch nicht gab, kamen zu ziemlich ähnlichen Ergebnissen [10].

Einschätzung

Unsere Einschätzung ist deshalb: Die Wirksamkeit medizinischer Interventionen ist schlechter belegt, als man meint. Bevor wir also handeln und uns für eine Intervention entscheiden, ob als Patient oder als Behandler, wäre es gut darüber nachzudenken, ob eine Intervention nötig und sinnvoll ist. Denn: Die Wirksamkeit ist vergleichsweise schwach belegt. Das Nebenwirkungspotenzial, vor allem von pharmakologischen Interventionen ist jedoch dort, wo es untersucht wird, größer als das Wirksamkeitspotenzial.

Das würde eigentlich nahelegen, den sog. Interventionsbias zu überdenken. Wir Menschen, vor allem Ärzte und Patienten, haben einen Interventionsbias, eine mentale Verzerrung dahingehend, dass wir denken, Eingreifen ist besser als Laufen lassen, Handeln ist besser als Nichtstun. Diese Verzerrung ist offensichtlich wenig gerechtfertigt. Es gibt sicherlich viele Fälle, vor allem akute, bei denen diese Haltung hilfreich ist. Aber es gibt offensichtlich auch viele, bei denen einmal länger Nachdenken oder Nachlesen und Abwarten die bessere Option ist. Jedenfalls wissen wir jetzt, dass ein bisschen mehr Skepsis im Umgang mit dem medizinischen Erlösungsnarrativ nicht nur angebracht und sachlich richtig ist, sondern eigentlich die aufgeklärtere und besser informierte Haltung.

Wisst Ihr jetzt, liebe Faktenchecker, Wissenschaftsredakteure und andere Medizinenthusiasten, warum ich den neuen, schlecht geprüften Impfplattformen gegenüber skeptisch bin? Das liegt an den Daten zu medizinischen Interventionen im Allgemeinen. Denn dadurch, dass wir hier eine Zufallsauswahl gezogen haben, lässt sich unser Ergebnis verallgemeinern.

Quellen und Literatur

1. McKeown T. Die Bedeutung der Medizin: Traum, Trugbild oder Nemesis? Frankfurt: Suhrkamp 1982; orig. 1976.
2. McKeown T. The Role of Medicine: Dream, Mirage, or Nemesis? London: The Nuffield Trust 1976.
3. Howick J, Koletsi D, Ioannidis JPA, et al. Most healthcare interventions tested in Cochrane Reviews not effective according to high quality evidence: a systematic review and meta-analysis. *Journal of Clinical Epidemiology* 2022;148 doi: https://doi.org/10.1016/j.jclinepi.2022.04.017
4. Howick J. The Philosophy of Evidence-Based Medicine. Chichester: Wiley-Blackwell 2011.
5. Storebø OJ, Pedersen N, Ramstad E, et al. Methylphenidate for attention deficit hyperactivity disorder (ADHD) in children and adolescents – assessment of adverse

events in non-randomised studies. *Cochrane Database of Systematic Reviews* 2018;5(CD012069) doi: https://doi.org/10.1002/14651858.CD012069.pub2
6. Storebø OJ, Ramstad E, Krogh HB, et al. Methylphenidate for children and adolescents with attention deficit hyperactivity disorder (ADHD). *The Cochrane database of systematic reviews* 2015;2015(11):Cd009885. doi: https://doi.org/10.1002/14651858.CD009885.pub2 [published Online First: 2015/11/26]
7. Schünemann H, Brozek J, Guyatt GH, et al. GRADE Handbook: Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. London: Cochrane Collaboration, 2013.
8. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology* 2011;64(4):401-06. doi: https://doi.org/10.1016/j.jclinepi.2010.07.015
9. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal* 2008;336:924. doi: doi: http://dx.doi.org/10.1136/bmj.39489.470347.AD
10. El Dib RP, Atallah AN, Andriolo RB. Mapping the Cochrane evidence for decision making in health care. *Journal of Evaluation in Clinical Practice* 2007;13:689-92.