Ein kurzer methodischer Kommentar anlässlich einer „Retraction“ unserer Homöopathie-ADHS-Meta-Analyse

Wir hatten uns zu früh gefreut. Ich hatte ja letzten Sommer berichtet, dass wir eine Meta-Analyse zu Homöopathie bei ADHS publizieren konnten, die eine signifikante Effektstärke von g = 0.6 zeigte [1]. Vor Kurzem wurde sie zurückgezogen („retracted“), und zwar vom Journal, nicht von uns.

Der Hintergrund dazu: Wir hatten einen Extraktionsfehler gemacht, und zwar eine Effektgröße positiv kodiert, die eigentlich negativ kodiert gehört. Das ist einer der Fallstricke in einer Meta-Analyse, über den ich jetzt selber gestolpert bin. Denn man muss sich immer fragen: Deuten nun die Effekte einer Studie in die Richtung der vermuteten Hypothese, unterstützen also die Vermutung, dass der Unterschied für die Wirksamkeit einer Behandlung spricht, oder dagegen? In diesem Falle [2] war das Ergebnis nicht nur nicht signifikant für die Homöopathie, sondern wies auch in die andere Richtung. Das hätte in der Analyse mit einem Minus-Zeichen versehen werden müssen, was ich schlicht und ergreifend übersehen hatte. Und meinen Kollegen ist es auch nicht aufgefallen und so hat sich dieser sehr dumme Fehler eingeschlichen.

Was ist der Effekt eines solchen Fehlers? In der original publizierten und jetzt zurückgezogenen Analyse ist der Effekt über alle sechs Studien g = 0.569, mit einer Irrtumswahrscheinlichkeit von p < .001. Das Schätzverfahren ist ein Modell zufälliger Effekte, das der weiten Streuung der Effektstärken gerecht wird. Der Effekt war also in dieser Analyse sehr deutlich. Mit dem korrigierten Vorzeichen ergibt sich ein g = 0.568 mit dem Modell zufälliger Effekte und ist damit von der Schätzung her sehr ähnlich. Was sich aber ändert, ist die Signifikanzschätzung. Sie verändert sich zu p = 0.053 und verpasst knapp die formale Signifikanzgrenze.

Wenn man nur die vier placebokontrollierten Studien betrachtet, ergibt sich neu eine Effektschätzung von g = 0.592, ebenfalls mit dem Modell zufälliger Effekte geschätzt. Die publizierte Analyse hatte g = 0.605 berichtet, also etwas höher. Dieser Effekt war signifikant mit p = 0.03. Jetzt ist der Effekt numerisch etwas kleiner; immer noch relativ groß, aber nicht mehr signifikant, nämlich p = 0.2. Mit einem Modell fester Effekte wäre

dieser Effekt kleiner ($g = 0.561$), aber signifikant ($p < .001$). Aber ein solches Modell wäre nicht angemessen, weil die Studien nicht homogen genug sind.

Wir sehen also: Das Vorzeichen hat vor allem einen Effekt auf die Signifikanz der Analyse, weniger auf die Schätzung der Größe des Effekts. Das liegt eben daran, dass der Effekt dieser Studie numerisch klein ist, im Vergleich zu den anderen Studien, vor allem im Vergleich zur Langzeitstudie aus Indien, die einen sehr großen Effekt von $g = 1.9$ aufweist und die Analyse dominiert. Daher ergibt sich nun durch das negative Vorzeichen dieser einen Studie eine viel größere Schwankungsbreite, die wiederum die Signifikanzschätzung beeinflusst.

Aufgrund dieser großen Schwankungsbreite ist auch ein Modell fester Effekte unangemessen, auch wenn es signifikante Effekte liefern würde.

Modell fester und zufälliger Effekte

Was ist der Unterschied? Bei einer Meta-Analyse legt man immer ein statistisches Modell an die Daten. Ein Modell fester Effekte geht davon aus, dass der wahre zu schätzende Effekt der Effekt des Mittelwertes aller Studien plus einer Schwankung, eines Stichprobenfehlers, ist. Diesen schätzt man aufgrund der Abweichungen der einzelnen Studien vom Mittelwert im Verhältnis zur Anzahl aller Studien, vergleichbar der Definition eines Standardfehlers in der normalen Statistik.

Das Modell zufälliger Effekte nimmt nun an, dass zusätzlich zum Stichprobenfehler noch eine systematische Schwankung dazukommt, deren wahre Größe man nicht kennt, sondern einfach schätzt, mit einem zusätzlichen Schätzverfahren. Man geht also davon aus, dass die wahren Werte nicht einfach um einen Mittelwert herum zufällig schwanken, sondern dass sie zufällig schwanken *und* dass noch eine systematische Schwankung hinzukommt. Das ist meistens die realistischere Annahme. Dieses Modell führt in der Regel, vor allem dann, wenn es angemessen ist, zu anderen, oft auch zu größeren Effektstärkeschätzungen, aber dafür zu konservativeren Signifikanzschätzungen. Denn die Signifikanz wird hier nicht nur aus dem Stichprobenfehler, sondern aus diesem und der systematischen Schwankung geschätzt.

In den Meta-Analysen, die ich bis jetzt gerechnet und gesehen habe, waren fast immer zufällige Effekte passend.

Die Retraction

Das Journal hat vor allem diesen Fehler moniert. Der war in der Tat ein Fehler. Wir hätten ihn gerne mit einem Korrigendum verbessert. Das wäre aus unserer Sicht möglich gewesen. Denn es ändert sich an der Gesamteinschätzung nicht viel. Diese war: Homöopathie ist vielversprechend, aber die Analyse gründet auf wenigen und zu weit streuenden Studien und daher sollte man das näher untersuchen. Was sich, wie ich gezeigt habe, ändert, ist weniger die Einschätzung der Größe des Effekts, sondern die Signifikanz des Gesamtmodells. Und in Sachen Signifikanz gibt es ohnehin sehr unterschiedliche Aussagen. Der Altmeister der psychologischen Methodenlehre, der Harvard-Methodiker Robert Rosenthal, hat mal einen Artikel publiziert, in dem er schrieb „Surely, God loves the 0.6 as he loves the 0.5“ [3, p. 1277]. Damit meinte er: Die Fixierung auf ein bestimmtes Irrtumswahrscheinlichkeitsniveau ist reine Konvention und nicht immer klug. Wichtig, das betonte er immer wieder und das hat sich mindestens in der Psychologie durchgesetzt, ist die Effektgröße selber. Dass man diese gegen Zufallsschwankungen absichern muss, versteht sich. Und so könnte man nun sagen: Die Effektgröße ändert sich nicht sehr, aber die Einschätzung, wie stark sie eine Zufallsschwankung repräsentiert, die ändert sich sehr wohl. Das stimmt. Aber das ändert nichts an unserer Einschätzung: Homöopathie bei ADHS ist auf jeden Fall interessant und sollte weiter untersucht werden. Im Übrigen ist mittlerweile eine neue Studie erschienen, die wir in eine verbesserte Analyse aufnehmen werden, die wir dann erneut publizieren werden, diesmal ohne Vorzeichenfehler.

Das Journal hat noch zwei weitere Punkte genannt: dass wir uns bei einer Risk-of-Bias-Einschätzung geirrt hätten und dass wir bei der Effektstärkeschätzung der indischen Studie die publizierten Effektstärken hätten verwenden müssen und nicht unsere eigene Schätzung. Zum letzten Vorwurf kann ich sagen: Das ist aus meiner Sicht falsch, weil die publizierten Effektstärkeschätzungen der indischen Publikation offensichtlich falsch waren. Warum ist eine andere Frage. Aber ich habe sie anhand der publizierten Daten nachgerechnet und sie sind falsch. Daher habe ich meine errechneten Effektstärken verwendet. Zur falschen Risk-of-Bias-Einschätzung kann ich nur sagen: Diese hängt sehr stark davon ab, welche Information man zugrunde legt. Oft publizieren Autoren nicht alles, was sie gemacht haben, z.B. weil ihnen nicht klar war, dass in 10 Jahren alle Leute nach dieser Information suchen würden und weil sie Platz sparen müssen. Wenn man aber weiß, wie die Autoren gearbeitet haben, weil man sie kennt und mit ihnen geredet hat, kann man andere Einschätzungen treffen. Man kann darüber streiten, ob das gut oder schlecht, möglich oder falsch ist. Außerdem sind manche Einschätzungen wirklich bis zu einem gewissen Grad sehr subjektiv. Man kann natürlich immer versuchen, auf die sehr konservative Seite zu schwenken. Wenn man das tut, dann ist nichts mehr wirklich gut und verlässlich, außer in sehr wenigen Fällen.

Der einzige aus meiner Sicht tatsächlich stichhaltige Fehler, den wir auch sofort zugestanden haben, war also der Kodierfehler. Ob man darauf mit einer Retraction reagieren muss, dieses Urteil überlasse ich anderen. Ich persönlich finde, man hätte auch mit einer Korrektur reagieren können.

Wenn ich zum Beispiel daran denke, dass die Arbeitsgruppe von Viola Priesemann eine Publikation in Science publiziert hat, die nachweislich und von ihr zugegebener Maßen mit falschen Daten operiert hat und diese Arbeit nicht zurückgezogen hat [4, 5], dann fragt man sich, mit welchen Maßstäben wer vermessen wird. Wir Homöopathieforscher, weil am Rande stehend, mit sehr scharfem Maß. Eine Arbeitsgruppe am Max-Planck-Institut, die das Lieblingsnarrativ der Regierung bedient, darf schon mal falsche Daten in ihr Modell einspeisen, ohne dass die FAZ nervös wird.

Wer mir das nicht glaubt: Wir haben das alles haarklein publiziert und mit Links nachgewiesen in unserer kürzlich publizierten Arbeit in *Futures* [6]. Dort ist auch der Blog von Frau Priesemann verlinkt, wo sie zugegeben hat, dass wir recht haben [5]. Ich kann den Artikel im PDF gerne allen schicken, die sich dafür interessieren. E-Mail genügt.

Was lerne ich draus? Ich werde garantiert keine Daten für Meta-Analysen nach 20 Uhr mehr kodieren.

Quellen und Literatur

1. Gaertner K, Teut M, Walach H. Is homeopathy effective for attention deficit and hyperactivity disorder? A meta-analysis. Pediatric Research. 2022; https://doi.org/10.1038/s41390-022-02127-3.
2. Jacobs J, Williams AL, Girard C, Njike VY, Katz D. Homeopathy for attention-deficit/hyperactivity disorder: a pilot randomized-controlled trial. Journal of Alternative and Complementary Medicine. 2005;11(5):799-806. Epub 2005/11/22. doi: https://doi.org/10.1089/acm.2005.11.799. PubMed PMID: 16296913.
3. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. American Psychologist. 1989;44:1276-84.
4. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. Science. 2020;369(6500):eabb9789. doi: https://doi.org/10.1126/science.abb9789.
5. Dehning J, Spitzner FP, Linden M, Mohr SB, Neto JP, Zierenberg J, et al. Model-based and model-free characterization of epidemic outbreaks – Technical notes on Dehning et al., Science, 2020. . Github. 2020. 6.
6. Kuhbandner C, Homburg S, Walach H, Hockertz S. Was Germany's Lockdown in

Spring 2020 Necessary? How bad data quality can turn a simulation into a dissimulation that shapes the future. Futures. 2022;135:102879. doi: https://doi.org/10.1016/j.futures.2021.102879.