Fallstricke der Meta-Analyse

Description

Ein kurzer methodischer Kommentar anlÃ¤sslich einer â??Retractionâ?? unserer HomÃ¶opathie-ADHS-Meta-Analyse

Wir hatten uns zu frÃ¼h gefreut. Ich hatte ja letzten Sommer berichtet, dass wir eine Meta-Analyse zu HomÃ¶opathie bei ADHS publizieren konnten, die eine signifikante EffektstÃ¤rke von g = 0.6 zeigte [1]. Vor Kurzem wurde sie zurÃ¼ckgezogen (â??retractedâ??), und zwar vom Journal, nicht von uns.

Der Hintergrund dazu: Wir hatten einen Extraktionsfehler gemacht, und zwar eine EffektgrÃ¶Ã?e positiv kodiert, die eigentlich negativ kodiert gehÃ¶rt. Das ist einer der Fallstricke in einer Meta-Analyse, Ã¼ber den ich jetzt selber gestolpert bin. Denn man muss sich immer fragen: Deuten nun die Effekte einer Studie in die Richtung der vermuteten Hypothese, unterstÃ¼tzen also die Vermutung, dass der Unterschied fÃ¼r die Wirksamkeit einer Behandlung spricht, oder dagegen? In diesem Falle [2] war das Ergebnis nicht nur nicht signifikant fÃ¼r die HomÃ¶opathie, sondern wies auch in die andere Richtung. Das hÃ¤tte in der Analyse mit einem Minus-Zeichen versehen werden mÃ¼ssen, was ich schlicht und ergreifend Ã¼bersehen 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 zurÃ¼ckgezogenen Analyse ist der Effekt Ã¼ber alle sechs Studien g = 0.569, mit einer Irrtumswahrscheinlichkeit von p < .001. Das SchÃ¤tzverfahren ist ein Modell zufÃ¤lliger Effekte, das der weiten Streuung der EffektstÃ¤rken gerecht wird. Der Effekt war also in dieser Analyse sehr deutlich. Mit dem korrigierten Vorzeichen ergibt sich ein g = 0.568 mit dem Modell zufÃ¤lliger Effekte und ist damit von der SchÃ¤tzung her sehr Ã¤hnlich. Was sich aber Ã¤ndert, ist die SignifikanzschÃ¤tzung. Sie verÃ¤ndert sich zu p = 0.053 und verpasst knapp die formale Signifikanzgrenze.

Wenn man nur die vier placebokontrollierten Studien betrachtet, ergibt sich neu eine EffektschÃ¤tzung von g = 0.592, ebenfalls mit dem Modell zufÃ¤lliger Effekte geschÃ¤tzt. Die publizierte Analyse hatte g = 0.605 berichtet, also etwas hÃ¶her. Dieser Effekt war signifikant mit p = 0.03. Jetzt ist der Effekt numerisch etwas kleiner; immer noch relativ groÃ?, aber nicht mehr signifikant, nÃ¤mlich p = 0.2. Mit einem Modell fester

Effekte wÃ¤re dieser Effekt kleiner (g = 0.561), aber signifikant (p < .001). Aber ein solches Modell wÃ¤re 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 SchÃ¤tzung der GrÃ¶Ã?e 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 groÃ?en Effekt von g = 1.9 aufweist und die Analyse dominiert. Daher ergibt sich nun durch das negative Vorzeichen dieser einen Studie eine viel grÃ¶Ã?ere Schwankungsbreite, die wiederum die SignifikanzschÃ¤tzung beeinflusst.

Aufgrund dieser groÃ?en Schwankungsbreite ist auch ein Modell fester Effekte unangemessen, auch wenn es signifikante Effekte liefern wÃ¼rde.

Modell fester und zufÃ¤lliger 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 schÃ¤tzende Effekt der Effekt des Mittelwertes aller Studien plus einer Schwankung, eines Stichprobenfehlers, ist. Diesen schÃ¤tzt man aufgrund der Abweichungen der einzelnen Studien vom Mittelwert im VerhÃ¤ltnis zur Anzahl aller Studien, vergleichbar der Definition eines Standardfehlers in der normalen Statistik.

Das Modell zufÃ¤lliger Effekte nimmt nun an, dass zusÃ¤tzlich zum Stichprobenfehler noch eine systematische Schwankung dazukommt, deren wahre GrÃ¶Ã?e man nicht kennt, sondern einfach schÃ¤tzt, mit einem zusÃ¤tzlichen SchÃ¤tzverfahren. Man geht also davon aus, dass die wahren Werte nicht einfach um einen Mittelwert herum zufÃ¤llig schwanken, sondern dass sie zufÃ¤llig schwanken *und* dass noch eine systematische Schwankung hinzukommt. Das ist meistens die realistischere Annahme. Dieses Modell fÃ¼hrt in der Regel, vor allem dann, wenn es angemessen ist, zu anderen, oft auch zu grÃ¶Ã?eren EffektstÃ¤rkeschÃ¤tzungen, aber dafÃ¼r zu konservativeren SignifikanzschÃ¤tzungen. Denn die Signifikanz wird hier nicht nur aus dem Stichprobenfehler, sondern aus diesem und der systematischen Schwankung geschÃ¤tzt.

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

Die Retraction

Das Journal hat vor allem diesen Fehler moniert. Der war in der Tat ein Fehler. Wir hÃ¤tten ihn gerne mit einem Korrigendum verbessert. Das wÃ¤re aus unserer Sicht mÃ¶glich gewesen. Denn es Ã¤ndert sich an der GesamteinschÃ¤tzung nicht viel. Diese war: HomÃ¶opathie ist vielversprechend, aber die Analyse grÃ¼ndet auf wenigen und zu weit streuenden Studien und daher sollte man das nÃ¤her untersuchen. Was sich, wie ich gezeigt habe, Ã¤ndert, ist weniger die EinschÃ¤tzung der GrÃ¶Ã?e 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 â??Surely, God loves the 0.6 as he loves the 0.5â?? [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 EffektgrÃ¶Ã?e selber. Dass man diese gegen Zufallsschwankungen absichern muss, versteht sich. Und so kÃ¶nnte man nun sagen: Die EffektgrÃ¶Ã?e Ã¤ndert sich nicht sehr, aber die EinschÃ¤tzung, wie stark sie eine Zufallsschwankung reprÃ¤sentiert, die Ã¤ndert sich sehr wohl. Das stimmt. Aber das Ã¤ndert nichts an unserer EinschÃ¤tzung:

HomÃ¶opathie bei ADHS ist auf jeden Fall interessant und sollte weiter untersucht werden. Im Ã?brigen 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-EinschÃ¤tzung geirrt hÃ¤tten und dass wir bei der EffektstÃ¤rkeschÃ¤tzung der indischen Studie die publizierten EffektstÃ¤rken hÃ¤tten verwenden mÃ¼ssen und nicht unsere eigene SchÃ¤tzung. Zum letzten Vorwurf kann ich sagen: Das ist aus meiner Sicht falsch, weil die publizierten EffektstÃ¤rkeschÃ¤tzungen 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 EffektstÃ¤rken verwendet. Zur falschen Risk-of-Bias-EinschÃ¤tzung kann ich nur sagen: Diese hÃ¤ngt 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 wÃ¼rden und weil sie Platz sparen mÃ¼ssen. Wenn man aber weiÃ?, wie die Autoren gearbeitet haben, weil man sie kennt und mit ihnen geredet hat, kann man andere EinschÃ¤tzungen treffen. Man kann darÃ¼ber streiten, ob das gut oder schlecht, mÃ¶glich oder falsch ist. AuÃ?erdem sind manche EinschÃ¤tzungen wirklich bis zu einem gewissen Grad sehr subjektiv. Man kann natÃ¼rlich immer versuchen, auf die sehr konservative Seite zu schwenken. Wenn man das tut, dann ist nichts mehr wirklich gut und verlÃ¤sslich, auÃ?er in sehr wenigen FÃ¤llen.

Der einzige aus meiner Sicht tatsÃ¤chlich stichhaltige Fehler, den wir auch sofort zugestanden haben, war also der Kodierfehler. Ob man darauf mit einer Retraction reagieren muss, dieses Urteil Ã¼berlasse ich anderen. Ich persÃ¶nlich finde, man hÃ¤tte auch mit einer Korrektur reagieren kÃ¶nnen.

Wenn ich zum Beispiel daran denke, dass die Arbeitsgruppe von Viola Priesemann eine Publikation in Science publiziert hat, die nachweislich und von ihr zugegebener MaÃ?en mit falschen Daten operiert hat und diese Arbeit nicht zurÃ¼ckgezogen hat [4, 5], dann fragt man sich, mit welchen MaÃ?stÃ¤ben wer vermessen wird. Wir HomÃ¶opathieforscher, weil am Rande stehend, mit sehr scharfem MaÃ?. 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 nervÃ¶s wird.

Wer mir das nicht glaubt: Wir haben das alles haarklein publiziert und mit Links nachgewiesen in unserer kÃ¼rzlich 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 dafÃ¼r interessieren. E-Mail genÃ¼gt.

Was lerne ich draus? Ich werde garantiert keine Daten fÃ¼r 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 â?? Technical notes on Dehning et al., Science, 2020. . Github. 2020. 6.
6. Kuhbandner C, Homburg S, Walach H, Hockertz S. Was Germanyâ??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.

Date Created
03.11.2023