A brief methodological commentary on the retraction of our homeopathy ADHD meta-analysis

We had rejoiced too soon. Last summer, I reported that we were able to publish a meta-analysis on homeopathy in ADHD, which showed a significant effect size of g = 0.6 [1]. It was recently retracted by the journal, not by us.

The background: We had made an extraction error, namely positively coding an effect size that should actually be negatively coded. This is one of the pitfalls in a meta-analysis that I have now stumbled across myself. Because you always have to ask yourself: Do the effects of a study point in the direction of the hypothesis, i.e. do they support the assumption that the difference speaks for the effectiveness of a treatment, or against it? In this case [2], the result was not only not significant in favor of homeopathy, but even pointed in the other direction. This should have been marked with a minus sign in the analysis, which I had simply overlooked. And my colleagues didn't notice it either, so this very stupid mistake crept in.

What is the effect of such an error? In the originally published and now retracted analysis, the effect across all six studies is g = 0.569, with an error probability of p < .001. The estimation procedure is a random effects model that does justice to the wide dispersion of the effect sizes. The effect was therefore very clear in this analysis. With the corrected sign, the result is a g = 0.568 with the random effects model and is therefore very similar in terms of the estimate. What changes, however, is the significance estimate. It changes to p = 0.053 and just misses the formal significance threshold.

If only the four placebo-controlled studies are considered, the new effect estimate is g = 0.592, also estimated with the random effects model. The published analysis had reported g = 0.605, i.e. slightly higher. This effect was significant with p = 0.03. Now the effect is numerically somewhat smaller; still relatively large, but no longer significant, namely p = 0.2. With a fixed effects model, this effect would be smaller (g = 0.561), but significant (p < .001). But such a model would not be appropriate because the studies are not homogeneous enough.

So we see that the negative sign primarily has an effect on the significance of the analysis, less on the estimate of the size of the effect. This is because the effect of this study is numerically small compared to the other studies, especially compared to the long-term

study from India, which shows a very large effect of g = 1.9 and dominates the analysis. Therefore, the negative sign of this one study now results in a much larger range of variation, which in turn influences the significance estimate.

Because of this large range of variation, a fixed effects model is also inappropriate, even if it would provide significant effects.

Fixed and random effects model

What is the difference? In a meta-analysis, a statistical model is always applied to the data. A fixed effects model assumes that the true effect to be estimated is the effect of the mean value of all studies plus a fluctuation, a sampling error. This is estimated based on the deviations of the individual studies from the mean in relation to the number of all studies, comparable to the definition of a standard error in normal statistics.

The random effects model now assumes that, in addition to the sampling error, there is also a systematic variation whose true size is not known, but is simply estimated using an additional estimation procedure. It is therefore assumed that the true values do not simply fluctuate randomly around a mean value, but that they fluctuate randomly *and* that there is also a systematic fluctuation. That is usually the more realistic assumption. This model usually leads to other, often larger effect size estimates, especially when appropriate, but to more conservative significance estimates. The reason is, that the significance is estimated not only from the sampling error, but also from the systematic variation.

In the meta-analyses that I have calculated and seen so far, random effects were almost always appropriate.

The retraction

The journal criticized this error in particular. It was indeed a mistake. We would have liked to have corrected it with a corrigendum. In our view, that would have been possible, because the mistake doesn’t change the overall assessment much. This was: homeopathy is promising, but the analysis is based on too few and too widely scattered studies and should therefore be examined more closely. What changes, as I have shown, is not so much the assessment of the size of the effect, but the significance of the overall model. And when it comes to significance, there are very different statements anyway. The old master of psychological methodology, the Harvard methodologist Robert Rosenthal, once published an article in which he wrote “Surely, God loves the 0.6 as he loves the 0.5” [3, p. 1277]. What he meant was that fixing on a certain level of probability of error is pure convention and not always wise. What is important, as he repeatedly emphasized and which has become

established at least in psychology, is the effect size itself. It goes without saying that this must be safeguarded against random fluctuations. And so you could say: the effect size doesn't change much, but the assessment of how strongly it represents a random fluctuation does change. That is true. But that doesn't change our assessment: homeopathy for ADHD is definitely interesting and should be investigated further. Incidentally, a new study has now been published, which we will include in an improved analysis, which we will then publish again, this time without sign errors.

The journal also mentioned two other points: that we had made a mistake in our risk-of-bias assessment and that we should have used the published effect sizes for the effect size estimate of the Indian study and not our own estimate. With regard to the last accusation, I can say that in my view this is wrong because the published effect size estimates of the Indian publication were obviously wrong. Why is another question. But I have recalculated them using the published data and they are wrong. Therefore, I used my own calculated effect sizes. All I can say about the incorrect risk-of-bias assessment is that it depends very much on what information you base it on. Authors often do not publish everything they have done, e.g. because they did not realize that in 10 years everyone would be looking for this information and because they need to save space. But if you know how the authors have worked, because you know them and have talked to them, you can make different assessments. You can argue about whether this is good or bad, possible or wrong. Moreover, some assessments are really very subjective to a certain extent. Of course, you can always try to err on the very conservative side. If you do that, then nothing is really good and reliable anymore, except in very few cases.

In my view, the only really valid error, which we also conceded immediately, was the coding error. Whether one has to react to this with a retraction, I leave that judgment to others. Personally, I think we could have reacted with a correction.

When I think, for example, that Viola Priesemann's working group published a paper in Science that demonstrably and admittedly operated with false data and did not retract this paper [4, 5], then one wonders what standards are applied to whom. Against us homeopathy researchers, because we are on the fringes, very strict standards are used. A working group at the Max Planck Institute, which serves the government's favorite narrative, is allowed to feed false data into its model without the FAZ getting nervous.

For those who don't believe me: We have published all of this in great detail and proven it with links in our recently published paper in *Futures* [6]. There is also a link to Ms. Priesemann's blog, where she admits that we are right [5]. I can send the article in PDF format to anyone who is interested. Just e-mail me.

What do I learn from this? I will definitely no longer code data for meta-analyses after 8 pm.

Sources and literature

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.