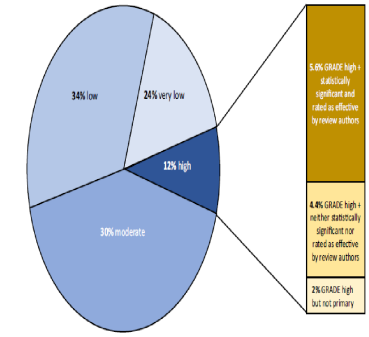

Fig. 1. Proportion of interventions according to their highest GRADE outcome (high, moderate, low, very low).

Meta-Review: The Backbone of Evidence Based Medicine Is Weak

Description

Only about 6 percent of all interventions used in medicine have sufficiently good data and are effective

Our new meta-review shows: The backbone of Evidence Based Medicine is weak

Regular readers of my texts know that I am very sceptical about the postmodern redemption narrative of modern medicine that proclaims: *We live so long and do so well because modern medicine has made such tremendous advances. Therefore, everything that modern pharmacology brings us is good, welcome and worthy of support (and should be funded by the public).*

Even the legendary social physician Thomas McKeown from Birmingham pointed out in the 1970s that this widespread popular opinion is most likely wrong and said in the introduction to his still very readable work “The Role of Medicine: Dream, Mirage, or Nemesis?” [1,2]: If he were St. Peter, he would only allow two types of doctors into heaven, namely trauma surgeons and dentists. Because those would be the only ones who really contributed to a reduction of suffering. The real progress and thus the extension of life span and the improvement of quality of life would not be due to medicine, but to socio-political progress, better nutrition, hygiene and living conditions without constant fear of hardship and death.

Well, that was in the 70s. Maybe it’s different today? We did a very large meta-study to answer the question of how good the data is for medical interventions in general. It has now been published in the Journal of Clinical Epidemiology [3]. I discuss the study and its findings in a little more detail in this blog. For those in a hurry: the data has not changed much. A maximum of 6 percent of all interventions used in medicine, no matter where, are covered by good data.

Only for a limited time (until 08-22-2022) the meta-review is freely accessible at: https://authors.elsevier.com/c/1fIHz3BcJQAobl

The study was initiated by Jeremy Howick, who worked for a long time at the Oxford Centre of Evidence Based Medicine and has worked very intensively on the conceptual penetration of Evidence Based Medicine (EBM) [4]. “Evidence” means “proof”, i.e. that which is supported by facts and data.

In this meta-review, our aim was to find out how many Cochrane Reviews really provide solid evidence of effectiveness, across all areas of medicine, from surgery to psychiatry, from paediatrics to gerontology and behavioural interventions.

The Cochrane Collaboration and Cochrane Reviews

The Cochrane Collaboration is originally a self-organised network of researchers, now a foundation. Researchers who work there do so out of scientific interest, and at most a small guild of principal investigators receives funding from local donors. The great hallmark of the Cochrane Collaboration has been independence from interest groups and industry funding. The Cochrane Collaboration is divided into so-called review groups, i.e. groups of authors who deal with certain larger areas, e.g. cardiology, oncology, etc., and within these with specific questions.

The Cochrane Collaboration publishes the so-called Cochrane Library, i.e. a collection of all reviews conducted by the researchers of the network. The website also lists protocols, i.e. definitions of reviews and their methodology, that are currently being carried out. And important reviews are regularly updated when new data become available.

The Cochrane Library thus represents the heart of EBM. This is because it summarizes the knowledge that is really important for individual specialist areas, diagnostic or treatment questions in medicine. This is done by summarizing all studies – especially randomized clinical trials, often also non-randomised cohort studies or case-control studies, depending on the research question – and evaluating them at the end. So if you want to know, for example, whether Ritalin works and is recommended for attention deficit hyperactivity disorder (ADHD) in children, you could search the Cochrane Library and find both a review on the effectiveness and one on the side effects of Ritalin [5,6]. In one case [6], a total of 38 studies involving more than 5,000 children were included. The authors concluded that all studies had design flaws that made them susceptible to bias. Therefore, it was unclear whether the small effect they found was really clinically significant. In the other case [5], 260 studies were included, demonstrating a high potential for side effects, but this is difficult to quantify.

These two examples are quite typical of Cochrane reviews; I mention them also because I talked about our own meta-analysis on homeopathy for ADHD in my last blog. The examples show how the authors proceed and how much synthesis work is involved.

Cochrane reviews are very rigid. They follow a pre-defined methodological grid and try to find and include as many studies as possible. To make the assessment easier, the so-called GRADE system was introduced around 2008: Grading of Recommendations, Assessments, Development and Evaluation, i.e. a kind of grading template [7-9]. In our own study, we have now examined one third of all Cochrane reviews that applied this GRADE system.

The selection of studies

Since we were only interested in the interventions where the GRADE system has been used to assess the data, i.e. since 2008 when it became common practice, the total data set of the Cochrane Library since then was 6,928

reviews large. If you took all the interventions into account, it would be even more. This is, of course, a huge amount of reviews that cannot be handled even by a large group of researchers. So we decided to randomly select a third of them and evaluate those; again, quite a lot of material. We were 12 researchers, and each of us thus had about 80 reviews to extract and sift through, on average. Of the 2,428 reviews selected and eligible, only 1,076 reviews met the inclusion criteria. These reviews covered a total of 1,567 interventions (because some reviews examine several interventions). The excluded reviews were mostly excluded because they did not contain a GRADE assessment or because they compared interventions with active controls. In fact, we only wanted those that compared an intervention with placebo, no treatment or standard therapy.

The procedure

We extracted each assigned number of studies into an Excel spreadsheet that had been previously tested. We were mainly interested in: is there an outcome in the review that captures the effect of the intervention, that was rated as “high quality” according to the GRADE system and, secondarily, is there evidence of side effects and harms and, if so, is there a GRADE indicator for that. So we extracted “high quality” outcomes and their effect size, but also side effects, and if such were documented also the “GRADE” indicator, besides some other variables of interest, e.g. whether placebo was controlled, standard treatment or no treatment, how many studies were in a review, which intervention, which population and which diagnosis had been studied.

GRADE

GRADE is a procedure that individually assesses the outcomes summarized in systematic reviews. If, for example, in lipid-lowering studies, blood lipid levels are recorded as a target criterion and only in some mortality studies (because blood lipid levels are of course much easier to record and more quickly available than mortality data), then GRADE would assess the outcome “blood lipid levels” as “low quality” or “very low quality” evidence. This is because these are surrogate parameters, and if the study was short or the documented effects small, then this outcome would influence the assessment of the outcome “blood lipids” as a whole, and the authors of such a review would perhaps write “low-quality evidence for effectiveness” or something similar. Conversely, mortality as an outcome or outcome parameter would be assessed as “high quality evidence” if it shows a clinically significant effect over a long period of time in a sufficient number of patients. This is because the GRADE system assesses not only significance, but also clinical meaningfulness, the numerical size of the effect and the appropriateness of the research design for the research question, as well as whether the patients studied were representative of the research question and whether the effect sizes were widely scattered. (More on this in the new meta-analysis chapter of my methodology blog.)

The GRADE system is therefore used to classify both the scientific and the clinical-practical significance of a finding. We were interested in the following: In how many reviews would we find clear indications that the individual studies on which the review is based have good data quality, so that the review would speak of “high quality of evidence”, i.e. a clear indication of clinical *and* scientifically proven effectiveness.

The result

Our random selection brought us reviews from all 53 Cochrane groups, i.e. across all clinical questions. Most interventions were tested in randomized trials. More than half of all interventions were pharmacological, 16% were psychological or behavioural, 6.4% were surgical, and other interventions such as diet and nutrition, exercise, alternative therapies each accounted for no more than 3% of the interventions. 45% of all interventions compared with placebo, 35% with standard therapy, and the rest with no treatment.

The result is best illustrated with our chart:

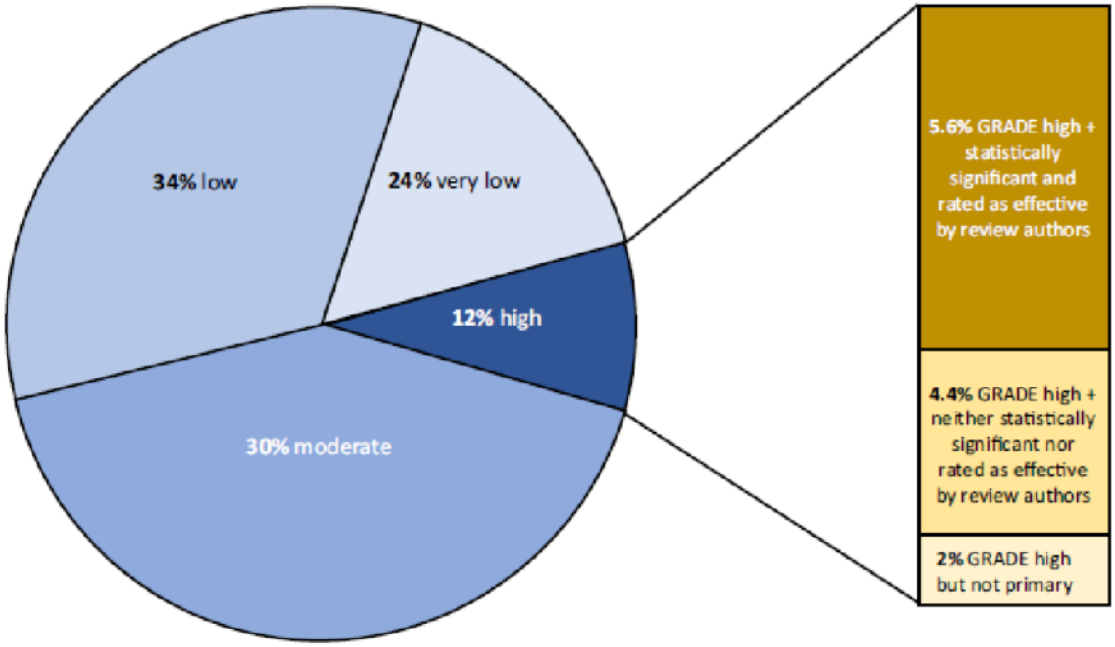

Fig. 1. Proportion of interventions according to their highest GRADE outcome (high, moderate, low, very low).

Figure from the original publication [3]: Proportion of interventions with high, moderate, low or very low GRADE rating

5.6% of all interventions had both an outcome with a “high quality” GRADE rating *and* a statistically significant effect (the dark brown bar). 4.4% of the reviews reported a GRADE rating that was classified as “high quality” but was neither significant nor rated as effective by the authors. 2% had a GRADE rating that was high, but only on a secondary outcome criterion. An example of this would be if an oncological study were to take the relapse-free time interval as the main objective criterion, but this is a surrogate parameter and would therefore not receive a “high quality” rating, and the quality of life as a secondary objective criterion, which would be “high quality” because it is clinically relevant, but was not the main objective parameter of the study.

You can see from the pie chart: In 58% of the reviews, the quality of the main objective criterion and thus of the clinical outcome is rated as “low” or “very low”, in 30% as moderate.

One can therefore state:

For less than 6% of all medical interventions, we have a high degree of clinical and scientific certainty that the intervention is effective and clinically useful.

This is, mind you, different from scientifically proven significance. This is because a study can generate significance that is nevertheless irrelevant for clinical application, e.g. because the effect is too small, because the target criterion was not clinically significant, because the study population was too specific so that the result is not transferable, and for a variety of other reasons.

Side effects

This modest efficacy profile contrasts with the side effects. Only 577 or 37% of all interventions also documented the side effects in such a way that they could be recorded in the reviews. Only a few, namely 6% of these studies, had an outcome for side effects that was described as “high quality”, e.g. mortality. 22% of all the reviews that documented side effects found significant harm, i.e. not just any side effects, but “harm”, i.e. harm caused by the intervention.

Influencing factors

We looked at possible study characteristics that could influence the outcome – diagnostic groups, study designs, intervention types, etc. – but found none.

Limitations

Even though we tried hard: No study is perfect. For example, one could argue that there are interventions that are clearly effective but did not fall into our grid because they are much older and do not need to be examined by studies done after 2008, splints of leg fractures, emergency surgery for arterial injuries or severe accidents, antibiotic therapy for bacterial pneumonia, insulin therapy for diabetes, gastric pumping for drug abuse, etc. This is certainly true, and to that extent our figure is perhaps a – slight – underestimate. Because earlier reviews, which also examined a large random sample of Cochrane reviews, without the GRADE assessment, which did not exist at the time, came to fairly similar results [10].

Assessment

Our assessment is, therefore: The effectiveness of medical interventions is less well documented than one might think. So before we act and decide on an intervention, whether as a patient or as a practitioner, it would be good to think about whether an intervention is necessary and clinically useful. Because: The effectiveness is comparatively weakly proven. However, the potential for side effects, especially of pharmacological interventions, is greater than the potential for efficacy where it is studied.

This would actually suggest rethinking the so-called intervention bias. We humans, especially doctors and patients, have an intervention bias, a mental bias to think that intervening is better than letting things happen, acting is better than doing nothing. This bias is obviously not very justified. There are certainly many cases, especially acute ones, where this attitude is helpful. But there are obviously also many where prolonged thinking or reading up and waiting is the better option. In any case, we now know that a little more scepticism in dealing with the medical redemption narrative is not only appropriate and factually correct, but actually the more enlightened and better informed attitude.

Do you now know, dear fact-checkers, science editors and other medical enthusiasts, why I am sceptical about the new, poorly vetted vaccination platforms? It’s because of the data on medical interventions in general. Because by drawing a random sample here, our result can be generalized.

Sources and literature

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.

Date Created
17.06.2022