Philosophy of Medicine

Original
Research
Article

On the Meaning of Medical Evidence Hierarchies

Jesper Jerkert
Department of Philosophy and History, Division of Philosophy, KTH Royal Institute of Technology, Stockholm, Sweden
Email: jerkert@kth.se

Abstract

Evidence hierarchies are investigative strategies ordered with regard to the claimed strength of evidence. They have been used for a couple of decades in EBM, particularly in assessing evidence for treatment recommendations, but remain controversial. An under-investigated question is what the order in the hierarchy means. Four interpretations are discussed here. The two most credible are "typically stronger" or "ideally stronger." The well-known GRADE framework seems to assume some "typically stronger" reading. Even if the interpretation of an evidence hierarchy were established, hierarchies are rather unhelpful for the task of evidence aggregation. Specifying the intended order relation may help to sort out disagreements.

ISSN 2692-3963 | Date Received 29-10-2020 | Date Accepted 16-03-2021 | Date First Published 19-04-2021
Philosophy of Medicine | DOI 10.5195/pom.2021.31 | Volume 2 | Issue 1 | pp. 1-21

1. Introduction

So-called evidence hierarchies (or hierarchies of evidence) have been used for a couple of decades, particularly in evidence-based medicine (EBM). In fact, the hierarchies seem to have emerged for use within clinical medicine (cf. Broadbent 2019, 139). They have then been tried out in related fields (for example, epidemiology) and can be found occasionally in areas such as public health. But evidence hierarchies have certainly not become established in science in general, and have probably never been adopted systematically in any natural science. On the contrary, even their use in the original setting of clinical medicine has become widely discussed and questioned. It is not difficult to understand why. An evidence hierarchy expresses fundamental tenets and provides methodological guidance for the task of assessing and weighing evidence from different sources. Assessing the strength of evidence for some hypothesis—for example, by assigning weights to different pieces of evidence from various sources—is a difficult task.

A general criticism against evidence hierarchies is that they are too schematic and do not account for the nuances needed in such assessments. There are two ways to respond to this critique. One is to say that evidence hierarchies are not intended to be used for advanced evidence weighing and assessment. Rather, the hierarchy is just a first approximation, a rule of thumb, or perhaps something to be shown to external parties to help them understand roughly how people think about evidence in EBM. The second possible answer is to acknowledge that evidence hierarchies are simple in their generic versions, but insist that they can be the basis of sophisticated evidence weighing and assessment, provided they are supplemented with various rules and instructions. The latter answer seems to be favoured by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group, currently the leading collaborative group on how to assess medical evidence in the EBM fashion. From 2011 to 2013, the GRADE working group produced a series of fifteen articles in the *Journal of Clinical Epidemiology* on detailed guidelines for evaluating medical evidence, with particular emphasis on how to move from a body of collected evidence to treatment recommendations (Guyatt, Oxman, Akl et al. 2011; Guyatt, Oxman, Kunz et al. 2011a, 2011b, 2011c, 2011d; Balshem et al. 2011; Guyatt, Oxman, Vist et al. 2011; Guyatt, Oxman, Montori et al. 2011; Guyatt, Oxman, Sultan et al. 2011, 2013; Brunetti et al. 2013; Guyatt, Oxman, Santesso et al. 2013; Guyatt, Thorlund, Oxman et al. 2013; Andrews, Guyatt, Oxman et al. 2013; Andrews, Schünemann, Oxman et al. 2013).[1]

If evidence hierarchies are supposed to be parts of advanced evidence assessment, then in order to build a systematic and coherent approach, we need to understand what the hierarchies represent in an epistemic sense: Whatever reasons are given for promoting a particular hierarchy, whenever some item is placed above another in a hierarchy, it is implied that the one above is "better," or is providing "stronger" evidence, than the one below. But how much stronger? Always stronger? Stronger only in certain circumstances? And even if these matters were settled, there is a follow-up question: what epistemically interesting features may follow from the claim that some item is above another? These are all aspects of one overarching question: what does the order in an evidence hierarchy mean? Even if evidence hierarchies are used only in a more hand-waving fashion (as rules of thumb, as pedagogical tools, or as something similar), it is still reasonable to ask what the order means: one item being above another surely has to mean something more precise than just that one is "stronger" than the other, even in a rule of thumb? So the question of

[1] These articles are still listed at the GRADE website (https://www.gradeworkinggroup.org) as recommended in-depth reading and thus represent the current state of the art of the GRADE approach.

the meaning of the order remains, however the importance and proper use of evidence hierarchies are judged in the end. This question has not been treated systematically in the literature so far. I will try to answer it.

In the next section, basic characteristics and terminology with respect to evidence hierarchies are discussed. Some ways in which the question of interest may change, thereby changing the design of the evidence hierarchy, are also reviewed. Section 3 contains the central contribution: an analysis of the order relations that are possible in an evidence hierarchy. Four interpretations of the order relation are distinguished, named, and discussed. Section 4 offers a more general discussion of the order relations detailed in the preceding section. The GRADE guidelines are included in the discussions of both sections 3 and 4. Section 5 briefly summarises and concludes. Throughout this text, I mainly focus on the use of medical evidence hierarchies for the overall purpose of recommending the best treatment.

2. Basic terminology and changes of question

The following is a fairly typical evidence hierarchy intended for use within EBM:

1. Systematic review of randomized trials or *n*-of-1 trials
2. Randomized trial or observational study with dramatic effect
3. Non-randomized controlled cohort/follow-up study
4. Case-series, case–control studies, or historically controlled studies
5. Mechanism-based reasoning. (OCEBM Levels of Evidence Working Group 2011)

Another example is the following:

1. N-of-1 clinical trial
2. Multiple-patient randomized trials
3. Observational studies: Patient-important outcomes
4. Basic research: Laboratory, animal, human physiology
5. Clinical experience. (Guyatt et al. 2015, 15)

Both of these hierarchies are ordered lists. Ordered lists of what? One could be inclined to say “study designs” or “study types,” but this does not seem entirely accurate. Mechanistic reasoning (“mechanism-based reasoning”) appears in the first hierarchy.[2] The item “clinical experience” appears in the second. None of these is a “study design” or a “study type,” according to the normal understanding of the terms. They can both be called *investigative strategies,* however, and this is fitting also for everything else found in the evidence hierarchies. Hence, in this article, I will refer to the items listed in an evidence hierarchy as “investigative strategies,” or just “strategies” for short. Instantiations of a particular strategy—for example, an actual randomized controlled trial (RCT), or an actual piece of mechanistic reasoning—is generically called an “investigation” in the present terminology. This word is preferable to the word “study” for the reason given above. However, I will use the phrase “observational study” since this is such an established term.

Since we will come back to the GRADE guidelines later, it is appropriate to look at the GRADE evidence hierarchy, too. In fact, the term “evidence hierarchy” (or “hierarchy of evidence”) does not appear in the 2011–2013 GRADE articles. The term “hierarchy of

[2] Mechanistic reasoning has appeared also in, for example, Straus et al. (2005, 169) in the form of “expert opinion…based on physiology, bench research or ‘first principles’ ”.

outcomes" appears once (Guyatt, Oxman, Kunz et al. 2011a, 399), and also appears in an earlier GRADE article (Guyatt et al. 2008, 995). Nonetheless, GRADE still contains an evidence hierarchy, though a very simple one, to be used in the process of determining which treatment is to be recommended. It has only two items: RCTs (above) and observational studies (below). Evidence from an RCT thus starts out above evidence from an observational study, but the evidence from either strategy can be further upgraded or downgraded, according to specified rules.

So, an evidence hierarchy is an ordered list of investigative strategies. The ordering, obviously, refers to the strength of evidence (or, at least, to the *claimed* strength of evidence).[3] This is the main point with designing an evidence hierarchy: to tell what strategies are better and which ones are worse at providing evidence for the issue at hand. The ordering is based on methodological distinctions and assessments: different strategies involve different methods, or method combinations, and those are claimed to have different advantages with respect to reliability, or to be susceptible to various unwanted biases. I will not discuss all existing methodological arguments for putting one strategy above another in a hierarchy (for example, for putting RCTs above observational studies). My project is rather this: given that it is claimed that one strategy should be above another strategy in a hierarchy, what could the order relation between those strategies possibly mean? I hope to show that interesting things can be said about this topic without having to deal with all arguments that have been proposed for putting one strategy above another.

One thing on which many people seem to agree is that evidence hierarchies are question-dependent: to the extent that a hierarchy is valid or useful at all, it is so in relation to some purpose or question. I will concentrate mainly on evidence hierarchies claimed to be useful for finding the best treatment against some disease or collection of symptoms. Even within this seemingly well-defined overall purpose, the exact question can change, which could trigger corresponding changes in the hierarchy. One such change has been discussed repeatedly in the literature. For a medical treatment to be recommendable, it ought to *both* cure the disease or ameliorate the symptoms *and* be (reasonably) free of adverse effects. Discussants have pointed out that the hierarchies will look rather different depending on whether you ask about maximizing the positive effects or about minimizing the negative effects of the treatment. Other ways of naming this contrast include efficacy versus risk (Osimani 2014, 298), benefits versus harms (Osimani and Mignini 2015, 1), or intended effects versus unintended effects (Vandenbroucke 2008a, 340–341, 2008b, 7; Osimani 2014, 295). Jan P. Vandenbroucke (2008a, 2008b) alternately phrased the difference as one between intended effects, on the one hand, and "discovery and explanation," on the other.

Although this distinction is not central to my main project, I believe a contribution to this debate can be offered. The terms that best describe what this particular distinction is about, I would argue, are *specified* versus *unspecified* outcome measures. In most cases, all of the proposed distinctions coincide in practice: the intended effects have been specified in advance and are precisely the beneficial effects that contribute to efficacy; and the non-intended effects are precisely the harmful side effects that increase risk and are unspecified. But they do not necessarily coincide, and when they do not, I believe that the distinction specified versus unspecified gets things right where the others fail. Consider an example

[3] A terminological note is that the term "strength of evidence," which I am using, is not used everywhere. The GRADE publications use "levels of evidence" in some places (for example, Balshem et al. 2011, 404). Even more common in the GRADE literature, however, is "quality of evidence," appearing in the titles, as well as *passim* in Balshem et al. (2011); Guyatt, Oxman and Kunz et al. (2011b, 2011c, 2011d); Guyatt, Oxman, Montori et al. (2011), inter alia.

where the medical condition is so serious that the most obvious intended and beneficial effect is survival. Rate of survival could then be taken as the corresponding outcome measure. It could be specified as such in advance and function as a proxy for efficacy. Since the medical condition is so serious, let us assume that a substantial fraction of patients will die unless they are treated. But the treatment available is also highly variable: some treated patients recover whereas others die. Furthermore, assume that it can be shown that at least a fraction of those who died, in spite of being treated, die from unwanted side effects of the treatment (in combination, naturally, with the underlying medical condition). It is then reasonable to say that an unintended effect of the treatment is death.[4] Death is also, obviously, a harm and a risk. The outcome measure for death is mortality. But here comes the point. Using the rate of survival or mortality as the main outcome measure will do equally fine, since they are, conceptually, the reversals of each other. Therefore, whatever evidence hierarchy is used for one is likely to be equally fitting for the other. This is at odds with claims that the hierarchy will change according to the distinctions of efficacy/risk, benefits/harms, or intended/unintended effects. The distinction specified versus unspecified outcome measure treats the example right: if rate of survival is a specified outcome measure, then so is mortality (since mortality is defined as the reversal of the rate of survival). Since both are specified to the same degree, then, it is not strange that the hierarchy remains unchanged whether we look at one or the other.

There are other question changes that may also change the hierarchy. One is about the population of interest. Consider this question: Which one of the two medical treatments T_A and T_B ought to be recommended for those who suffer from the given disease *D*? Here, the population of interest is everyone who suffers from *D*; that is, the question is about the effects of T_A and T_B averaged over many patients. But another question is: Which one of the two medical treatments T_A and T_B ought to be recommended for this patient, who suffers from *D*? Here, the target population is a single specified patient. The distinction between target populations seems to account for the fact that some published evidence hierarchies claimed to be valid for beneficial treatment effects put RCTs, or aggregations of RCTs, at the top (for example, Straus et al. 2005, 169), whereas others put so-called *n*-of-1 trials at the top (for example, Guyatt et al. 2015, 15). It would make no sense to have *n*-of-1 trials at the top if one is interested in average effects over patient populations, but it does make sense for the very patient who would participate (or has participated) in an *n*-of-1 trial.

Moreover, certain health conditions—or certain ways of measuring health conditions—are sensitive to psychological factors, which may, in turn, be influenced by the investigative strategy chosen. Thus, to the extent that the health outcome of interest is affected by human behaviour (and by human intentions, emotions, and so on), there will be situations where experimental control conditions negatively affect the reliability of measuring the outcome, as the control conditions may make people behave (feel, wish, and so on) significantly differently from a situation where there were no (or less) such conditions imposed. For example, an observational (but non-experimental) study may be claimed to provide stronger evidence than an experiment (for example, an RCT) when experimental control conditions are judged to have substantially negative effects on the reliability, although the reversal would have been true when other outcome measures are used. The appropriate design of the hierarchy could thus be dependent on the type of outcome measure chosen.

[4] Perhaps one might argue that death should not count as an "effect of the treatment" but is rather an effect of the non-intended side effects of the treatment. But I take "effect" to cover both wanted and unwanted outcomes. If death is due to the side effects of the treatment, then it is reasonable to say that death is an effect of the treatment.

I do not intend to list all the ways in which a change in the question could lead to a change in the evidence hierarchy. One more thing has to be mentioned, though. An evidence hierarchy could be appealed to in rather different "decision contexts." I am thinking mainly of differences in the available time and resources. One context is that of a single clinician who is to decide on the treatment (or diagnosis, and so on) for a patient seeking immediate help or advice. The clinician may then have just a couple of minutes to reach a decision, but even in such a short period of time, it would be possible to perform a literature search and to skim through a few research papers in order to make a more informed decision. In fact, in many early EBM publications, it was explicitly stated that the EBM way of thinking, including appeals to its evidence hierarchies, could and should be used in the everyday work of clinicians.[5] The assessment and recommendation rules promoted by GRADE, on the other hand, are clearly intended to be used by professional bodies and agencies, where a team of researches may be given months or years to reach a verdict.[6] The latter decision context is very different from that of a single clinician under time pressure. It is conceivable that a hierarchy judged to be fitting for one context is unfitting for the other. More subtly, it is possible that a particular hierarchy is judged to be correct for both settings (that is, the same list of strategies, in the same order) but the order *interpretations* of the hierarchy differ between the settings. Since this article is about the interpretations of order relations, this is an interesting idea. However, I am primarily interested in the decision contexts of professional bodies working with plenty of resources in terms of personnel, time, and money. One reason for being less interested in the busy clinician context is that busy clinicians have got, over the last couple of decades, increased access to up-to-date consensus documents on treatment recommendations, which should facilitate the decisions under time pressure, and with the help of which the need for a practising clinician to engage directly with original research literature, and therefore with evidence hierarchies, is downplayed. Those consensus documents have often been produced precisely by teams of researchers that work in the EBM way (for example, using GRADE), but have much more time and resources than the single clinician.[7]

3. The order relation

What *could* the order relation in an evidence hierarchy mean, and can we find some interpretations of the meaning more credible than others in our context of interest? To answer these questions, we need to find out what order interpretations there are, and then discuss which ones are reasonable and which ones are less so. There seem to be four main interpretations of the order relation available. Unfortunately, to the extent that they have been distinguished at all in the EBM context, they have been named in various ways. I will therefore introduce my own terminology, which differs from those of earlier authors. In discussing the four interpretations, I will assume that strengths of evidence are available on

[5] The idea is expressed clearly in the "founding paper" of EBM (Evidence-Based Medicine Working Group 1992) and also in a number of EBM publications that have been presented as handbooks, implying use in clinical reality (for example, Straus et al. 2005; Guyatt and Rennie 2002; Nordenstrom 2007).

[6] There is a list on the GRADE website with over 100 organizations and agencies that have publicly stated that they rely on the GRADE guidelines in their evidence assessment work. https://www.gradeworkinggroup.org.

[7] It seems that differences between decision contexts could have been highlighted more in some literature on evidence hierarchies. For example, Osimani (2014, 295) says clearly at the outset of her article that evidence hierarchies "are intended as a decision heuristics for professionals overwhelmed by data of heterogeneous quality and pressed by time constraints." Still, her article extensively discusses Vandenbroucke's (2008a, 2008b) ideas about hierarchies, although Vandenbroucke consistently discusses different ways of performing *research* and never explicitly considers a context in which a single clinician is supposed to make an everyday decision about a patient.

a continuous scale for each strategy. In other words, it is not the case that the strength of evidence may only assume one or a number of specific point values for any strategy, but rather a range of possible values. I will also assume that the strengths of evidence associated with investigations using a particular strategy can be represented graphically in the manner of probability distributions, with strength of evidence being the independent variable (x axis, increasing from left to right) and the probability being the dependent (y axis). One can think of these graphs either as real probability distributions, or more loosely as idealized empirical distributions of the evidential strengths of performed or imagined investigations. Example distributions are shown in Figure 1. Each distribution has a maximum and a minimum strength of evidence along the x axis. (Thus, they are not like Gaussians that stretch indefinitely to the left and to the right in the graphs.) All drawn distributions will be unimodal, but unimodality is important only for one specific interpretation (to be clarified later).

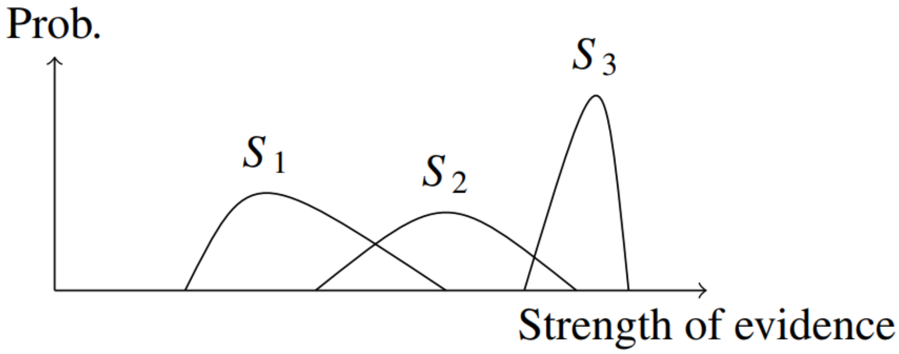

Figure 1. Example distributions of the strengths of evidence for the three strategies S_1, S_2, S_3. The vertical axis, labelled "Prob.", represents probability or frequency. As seen, no assumptions are made about the distributions being symmetric or asymmetric.

A reader familiar with GRADE may perhaps wonder whether the assumption of a continuous scale for strength of evidence is compatible with GRADE, where a discrete rating scale is used. It is. The GRADE working group acknowledges that "the quality of evidence represents a continuum" (Balshem et al. 2011, 404). The discrete GRADE scale with four possible values for the strength of evidence (high, moderate, low, very low) is thus a simplification of what is really a continuous scale.

3.1. **Stronger means** NON-OVERLAPPING STRONGER

Let us say that we are comparing investigative strategies S_1 and S_2, where S_1 is above S_2 in the hierarchy. This means that an investigation using strategy S_1 provides stronger evidence than an investigation using strategy S_2. A very strict interpretation of how to view, more precisely, the strengths of evidence from two investigations would be the following:

> NON-OVERLAPPING STRONGER
> Any S_1 investigation provides stronger evidence than any S_2 investigation.

In a graphical sense, this interpretation corresponds to non-overlapping ranges of evidential strengths for the S_1 and S_2 investigations, respectively; hence, the suggested name NON-OVERLAPPING STRONGER. This is shown in Figure 2. The interpretation implies that even though there may be differences regarding the strength of evidence among individual investigations using either of the two strategies, any single investigation using strategy S_1 is guaranteed to be evidentially superior to any single investigation using strategy S_2.

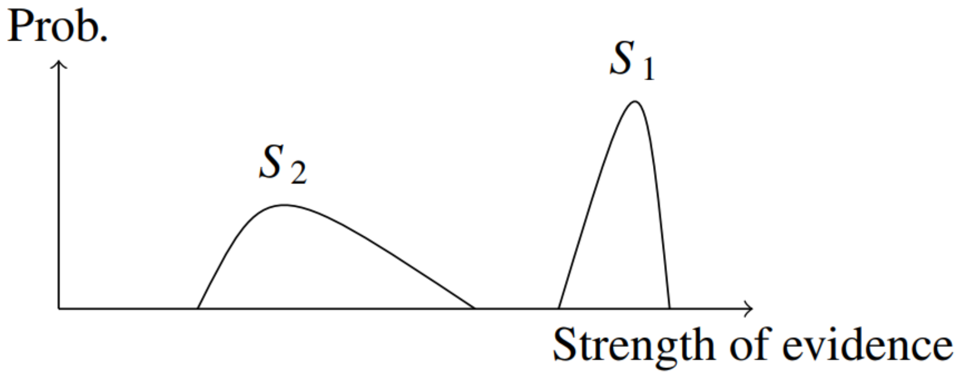

Figure 2. The NON-OVERLAPPING STRONGER interpretation entails that any investigation using strategy S_1 provides stronger evidence than any investigation using strategy S_2. This corresponds graphically to the fact that the distributions S_1 and S_2 do not overlap anywhere.

This interpretation is reminiscent of what mathematicians call "lexicographic ordering." This is an ordering principle according to which factors are ranked, and the value pertaining to the most important factor decides the sorting, irrespective of the values of the lower-ranked factors. Only if the values of the most important factor are equal, you look at the second most important factor, and so on. This is the sorting principle used in dictionaries (alphabetical order), hence the name. In a dictionary, the first letter of the word is the most important. "Apple" comes before "banana" because *a* comes before *b*, no matter what other letters follow. The analogy with evidence hierarchies is that the investigative strategy is the most important factor, thus acting as the "first letter." For example, it could be claimed that an RCT is above an observational study, no matter what other things can be said about the investigations, according to this interpretation. However, the analogy with the alphabetical ordering of words seems to break down after the first letter, as I am not aware of anyone having suggested, in the medical context, that some specified factor is the second-most important (the "second letter"), yet another is the third-most important factor (the "third letter"), and so on. Depending on the chosen definition, this might still be called "lexicographic ordering," albeit only in a rudimentary or partial guise.[8]

An even stronger variant of NON-OVERLAPPING STRONGER is readily available, although I will not assign a separate name to it:

> Any single S_1 investigation provides stronger evidence than any number of S_2 investigations combined.

According to this interpretation, then, no matter how many S_2 investigations are being performed, their aggregated evidence can never trump the evidence provided by an S_1 investigation.

[8] As for the history of the term "lexicographic order(ing)" in the principled, mathematical sense (not just in the context of words in dictionaries), it has been used for centuries. It is mentioned, for example, in both of Carl Friedrich Hindenburg's two compendiums of combinatorics (Hindenburg 1796, 201–202, 280; Hindenburg 1800, 15, 17). Hindenburg also says (1800, 106, 107) that the idea was described by Abraham de Moivre (1667–1754), but I have not been able to verify this claim.

Is the NON-OVERLAPPING STRONGER interpretation reasonable in the context of assessing evidence for recommending medical treatments? Hardly. The common understanding of different investigative strategies (for example, RCTs versus observational studies) is that they can be more or less credible, more or less prone to bias, more or less well performed. All of this potential variability makes it difficult to argue conclusively that the ranges of evidential strengths cannot overlap at all. If S_1 is considered evidentially superior to S_2 and the evidential ranges of S_1 and S_2 cannot overlap, then there is at least one methodological difference between S_1 and S_2, which accounts for the impossibility of an overlap. Although candidates have been discussed (particularly the presence or absence of randomization), I am not aware of a convincing argument to the effect that such a methodological difference between strategies justifies the NON-OVERLAPPING STRONGER interpretation. Also, as I argued in Section 2, the right design of a hierarchy may be dependent on the type of outcome measure chosen. This contributes to the difficulty of placing one strategy above another in the definitive way of the NON-OVERLAPPING STRONGER interpretation.

Nonetheless, in the EBM context the NON-OVERLAPPING STRONGER interpretation, or what appears to be it, has been claimed to be the correct one (or, at least, to be a plausible one) on many occasions, among which we may sample the following five from critics of various aspects of EBM:

- Adam La Caze (2008, 357) writes: “The EBM handbooks suggest a categorical interpretation of the hierarchy.” With respect to the claim that randomization is categorically better than non-randomization, he offers the following explanation: “*All* the results of a randomized study are *always* superior to the results of studies from lower down the hierarchy” (2008, 358). I am not sure that I understand exactly what the quoted sentence means, but something identical or similar to the NON-OVERLAPPING STRONGER interpretation seems to be intended.
- John Concato (2004) presents arguments against “a rigid hierarchy.” He seems to have the NON-OVERLAPPING STRONGER interpretation in mind, judging from the following formulation: “The conventional wisdom suggests that observational studies consistently provide biased results compared with randomized, controlled trials, regardless of the type of observational study or how well it was conducted” (2004, 342).
- Jason Grossman and Fiona J. MacKenzie (2005, 517) say that EBM supporters claim that “any study using the RCT design is considered superior to any study not using this design—and nowhere is this demonstrated more clearly than in EBM’s method of evidence evaluation, whose evidence hierarchies feature RCTs above all other forms of evidence.” This is definitely the NON-OVERLAPPING STRONGER interpretation.
- Robyn Bluhm (2009, 259n.) writes: “Note that a single RCT outweighs a review of observational studies, no matter how many or how good. Although in practice EBM advocates would likely agree that a well-designed nonrandomized study should outweigh a poorly designed RCT, in principle the hierarchy of evidence does not allow this judgment.”
- Alex Broadbent (2019, 135) says that “EBM arranges types of evidence into a pyramid” and asserts that “the pyramid is a lexical ordering, with higher kinds of evidence trumping lower kinds, meaning that very small quantities of higher-level evidence can outweigh very large quantities of lower kinds.”

Some of the criticisms against EBM proponents for embracing what appears to be the NON-OVERLAPPING STRONGER interpretation could be unfair, however. Certainly, there are quotations from EBM supporters that suggest that the NON-OVERLAPPING STRONGER interpretation is what they support. But you can also find quotations from EBM supporters that run counter to this interpretation. For example, Jorgen Nordenstrom (2007, 40) writes: “The hierarchy is not absolute, however. If, for instance, there is a large, clear-cut, therapeutic effect, the value of an observational study may be higher than that of many RCTs.” And others have concurred (for example, Guyatt and Rennie 2002, 7; Howick 2011, 56). It seems that EBM supporters—at least some of them—have gradually recognized that the NON-OVERLAPPING STRONGER interpretation of the order relation is untenable.[9]

But this shift is likely to be connected also to a shift in interest with respect to the “decision context” (in the sense indicated in Section 2). The decision context of greatest initial interest in EBM was that of a single clinician seeing a patient and having to reach a decision within a couple of minutes. Here, the NON-OVERLAPPING STRONGER interpretation may be adequate. The definitive (and later much-criticized) advice, “if the study wasn’t randomized, we’d suggest that you stop reading it and go on to the next article in your search” (Straus et al. 2005, 118), may then be reasonable in such a decision context. But later on, interest within EBM has turned more to the decision context in which a group of researchers have plenty of time to assess the evidence in some medical matter. The latter context is where my main interest lies, too. In that context, the NON-OVERLAPPING STRONGER interpretation is untenable.[10]

The GRADE system is not compatible with the NON-OVERLAPPING STRONGER interpretation. Whereas RCTs start out above observational studies, subsequent upgrading and downgrading may flip the ultimate order.[11]

3.2. **Stronger means STRONGER *CETERIS PARIBUS***

Let us see how we can make the interpretation weaker. Obviously, this requires that we drop the condition that the evidential ranges must not overlap. Once overlaps are permitted, however, there is still room to add different constraints that would justify S_1 being above S_2 in the list. One would be to say that an investigation using S_1 always yields stronger evidence than an investigation using S_2 if they are “alike” or “comparable” in all relevant respects except for the fact that they instantiate different strategies. This is a *ceteris paribus* condition, so we may formulate the interpretation as follows:

> STRONGER CP
> Any investigation using S_1 provides stronger evidence than any investigation using S_2 *ceteris paribus*.

Graphically, this means that we are able to map the evidential ranges of the investigations using S_1 and S_2, respectively, onto one another; and for each strength associated with an

[9] A special reason for the belief that a strong interpretation was originally intended in the context of EBM is the very word “hierarchy,” which, one could argue, by itself points in the direction of the NON-OVERLAPPING STRONGER interpretation.

[10] As for the EBM critics quoted above as suggesting that the NON-OVERLAPPING STRONGER interpretation is the intended one in EBM, two actually seem to think of the single clinician context (La Caze 2008; Concato 2004), whereas one alludes to both the single clinician and the research contexts (Broadbent 2019). In the two remaining references (Grossman and MacKenzie 2005; Bluhm 2009), only the research context seems to be considered.

[11] To be precise, according to the GRADE system *outcomes* are graded, rather than *investigations*. From a single investigation, different outcomes may therefore receive different final grades (Guyatt, Oxman, Akl et al. 2011, 385). Nonetheless, a generally applicable rule in GRADE says that evidence from an RCT starts out above evidence from an observational study *for all outcomes*. This is why it is reasonable to say that the RCT strategy is ranked above the observational study strategy in GRADE.

investigation using S_1, the corresponding strength (or strengths) of the *ceteris paribus* corresponding investigation (or investigations) using S_2 is lower. I mention both a singular and a plural variant, since, in principle, it would be possible for an investigation using S_1 to have more than one *ceteris paribus* corresponding investigation using S_2, and these S_2 investigations could provide non-identical strengths of evidence. Such a one-to-several mapping would result if there is some variable property that is completely irrelevant for the strength of evidence under S_1, but whose value is relevant for the strength of evidence under S_2. Different values of this property could then lead to different overall strengths of evidence under S_2. An illustration of the interpretation in a not too complicated case is given in Figure 3.

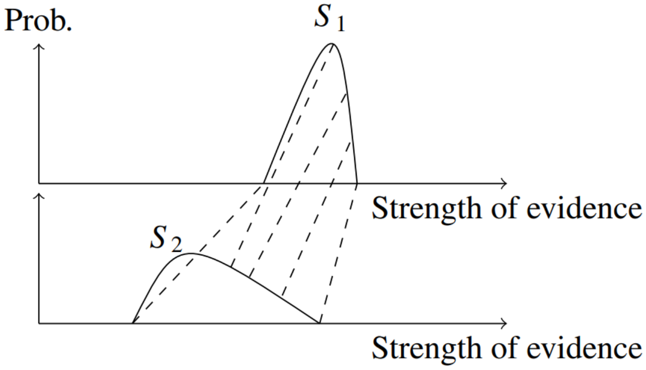

Figure 3. Illustration of the STRONGER CP interpretation. Strategy S_1 (top) provides stronger evidence than strategy S_2 (bottom) according to this interpretation if, for each strength of evidence under S_1, the *ceteris paribus* corresponding strength (or strengths) of evidence under S_2 is (are) lower. The *ceteris paribus* correspondence is here illustrated with dotted lines. No matter how complicated the correspondences may be with the possibility of one-to-several mappings of evidential strengths, the crucial property of these lines is that they must all slope in the same direction for the STRONGER CP interpretation to hold water. In fact, the shapes of the two distribution curves are irrelevant here.

In the context of EBM, the STRONGER CP interpretation seems to have been mentioned by Jacob Stegenga (2014), who claims, first, that no one (any longer) makes an "absolute" interpretation—by which he most probably means a NON-OVERLAPPING STRONGER interpretation. A weaker interpretation, which he dubs "categorical," is therefore more plausible, according to Stegenga. The interpretation "holds that the ordering of methods is categorical…that tokens of methods from higher on an evidence hierarchy are *necessarily* more reliable, *ceteris paribus*" (Stegenga 2014, 314). I take it that the STRONGER CP interpretation is what Stegenga has in mind here.[12]

Is the STRONGER CP interpretation more plausible in the context of EBM, as suggested by Stegenga? For a comparison between, for example, an RCT and mechanistic reasoning, I fail to see that this is the case: these two investigative strategies are so different from each other that it is difficult to understand what a *ceteris paribus* condition would even mean. This seems to imply either that the STRONGER CP interpretation is wrong, or that mechanistic reasoning should not be included in the hierarchy. For the sake of argument, let us also see whether the STRONGER CP interpretation makes sense in the well-known comparison between RCTs and observational studies.

[12] Stegenga is very critical of the use of evidence hierarchies at all. The fact that he seems to suggest a particular interpretation of the order relation in a hierarchy does not imply that he endorses hierarchies overall.

The necessary differences between RCTs and observational studies are that in an RCT there are experimental intervention and randomization, whereas in an observational study there is neither.[13] In all other respects, RCTs and observational studies *could* be alike. And there are certainly examples of observational studies being designed so as to become comparable to RCTs in various aspects—for example, cohort studies using propensity score matching. Nevertheless, many observational studies are not so similar to RCTs. The latter typically have stringent inclusion and exclusion criteria, whereas observational studies typically do not. Another way of putting this is that RCTs are "clean trials" but observational studies are not (Bluhm 2009, 259). This makes it very hard to understand what a *ceteris paribus* condition would mean in a comparison between actually occurring RCTs and observational studies (cf. Grossman and MacKenzie 2005, 520). In sum, *pace* Stegenga, it is difficult to make sense of the STRONGER CP interpretation in the context of EBM.[14]

3.3. **Stronger means TYPICALLY STRONGER**

It seems, then, that in the context of EBM we must drop the *ceteris paribus* interpretation of the order relation. To be able to retain the order and justify strategy S_1 being above S_2 in the list, we need to impose some other (weaker) condition on what the different positions of S_1 and S_2 mean. There are two principled ways of doing this. Either one compares some points where the weight lies in the distributions, or one compares some extreme values (maxima or minima) of the distributions. The latter option will be discussed in the next subsection. The former option (looking where the weight lies) gives something like this:

> TYPICALLY STRONGER
> An investigation using S_1 typically provides stronger evidence than an investigation using S_2.

The word "typically" can be read in various ways. One specification of what it means could be as follows: given the distributions of the strengths of evidence for investigations using S_1 or S_2, respectively, when we consider the most frequently occurring (the most probable) strength for each strategy, we will find that it is greater for investigations using S_1 than for investigations using S_2. Obviously, this corresponds to comparing the *modes* (by definition, the most frequently occurring value) of the respective probability distributions. The distributions are required to be unimodal for this interpretation to make sense. Other readings of "typically" are also possible. We could consider, instead of the mode, some other average-like function:

- The *expectation* of the strength of evidence for an investigation using S_1 is greater than the expectation of the strength of evidence for an investigation using S_2.
- The *median* strength of evidence for an investigation using S_1 is greater than the median strength of evidence for an investigation using S_2.

The formulation with expectations requires that strength of evidence is measured on (at least) an interval scale. There is hardly a consensus in science—or, more specifically, in

[13] Everyone agrees that in an observational study there is no randomization. Everyone ought also to agree that in an observational study there is no *experimental* intervention, as an observational study is no experiment. However, if you change the formulation to "intervention influenced or controlled by someone other than the patient," then things get more complicated. The basic question is how to define "observational study"; see Gugiu and Gugiu (2010, 239–240). This question does not matter for the overall argument in the main text, however.
[14] It is possible that I am misinterpreting Stegenga: rather than the STRONGER CP interpretation, he may have the IDEALLY STRONGER interpretation in mind, as he talks about a hierarchy being "constituted by ideal *types* of evidence" (2014, 315). The latter interpretation is discussed below.

medicine—on how to measure strength of evidence, and even less consensus that the right scale type is interval scale. The most common scale type seems rather to be the ordinal scale, with ordered labels such as strong/moderately strong/weak (as in GRADE), or with some numerical but still ordinal grading system such as the Jadad score (Jadad et al. 1996). So, the interval scale assumption cannot be taken for granted. This leaves the mode version and the median version. Both will do in principle. One advantage of the median version is that unimodality is not required, whereas it is required for the mode version. Each seems compatible with the GRADE framework, but nowhere in the GRADE papers is there an explicit acknowledgement that the TYPICALLY STRONGER interpretation is the correct one for the initial placement of RCTs above observational studies. Figure 4 illustrates the TYPICALLY STRONGER interpretation in the mode version.

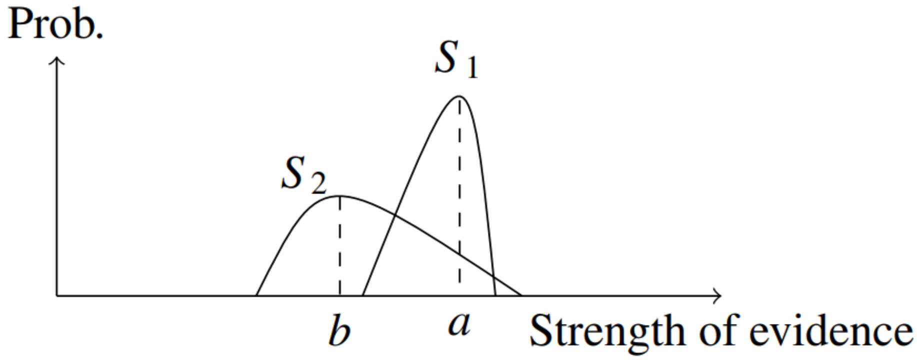

Figure 4. Illustration of the TYPICALLY STRONGER interpretation. S_1 provides stronger evidence than S_2 since the mode of S_1, labelled a, is larger than the mode of S_2, labelled b. (In this case, the verdict would have been the same for the medians, but it would have been possible to move these two differently skewed distributions towards one another in such a way that the median of S_1 would have been smaller than the median of S_2 even though mode a is still greater than mode b.) The minimum and maximum ranges (strengths of evidence) of the distributions are irrelevant under this interpretation.

3.4. Stronger means IDEALLY STRONGER

If we do not look at points in the distributions where the weight lies, the remaining option is to look at or near the extremes. One alternative would then be to consider the minima along the x (strength of evidence) axis, and to judge as the superior investigative strategy that which has the largest minimum. This rule would correspond to following a maximin principle (maximizing the minimum). However, there are two reasons for being sceptical about adopting the maximin principle in our context of interest. First: how bad can an investigation be? This is not so easy to tell in a coherent manner. For instance, what would the worst possible RCT look like? Could it, for example, be judged not to be randomized? Probably not, as an RCT is supposed to be randomized by definition. But then there is a problem of telling how sloppy and inadequate randomization measures can be without ceasing to count as randomization at all. Generally, the problem is one of anchoring the worst possible investigation of a particular strategy in well-defined methodological features while still claiming that the investigation is instantiating the strategy. I take this to be a difficult task. Secondly: even if we could understand what constitutes the methodologically worst investigations of each strategy, why would we be interested in these inferior investigations? Do we believe that they are common? Why, in spite of being bad, are they so important that we build our system of ranking strategies on them? Also, from the perspective of promoting good investigations in the future (to the extent that this is an appropriate use of evidence hierarchies at all), the following would not be the most reassuring and positive message to convey: We believe that more investigations using strategy S_1 are needed because even if the worst of them are really bad, they are still better than the worst investigations using strategy S_2.

I would argue, then, that to the extent that we should look at extremes, it is more appropriate to compare the highest evidential strengths available between strategies. The most straightforward interpretation of why S_1 is above S_2 would then be:

> IDEALLY STRONGER
> The ideally performed S_1 investigation provides stronger evidence than the ideally performed S_2 investigation.

Another way of phrasing the interpretation is this: there is a possible S_1 investigation that necessarily provides stronger evidence than any S_2 investigation. The interpretation is illustrated in Figure 5, but note that S_1 and S_2 have been swapped: the graph shows S_2 as ranked above S_1. The reason for this is that the distributions happen to be identical to those in Figure 4, facilitating a comparison.

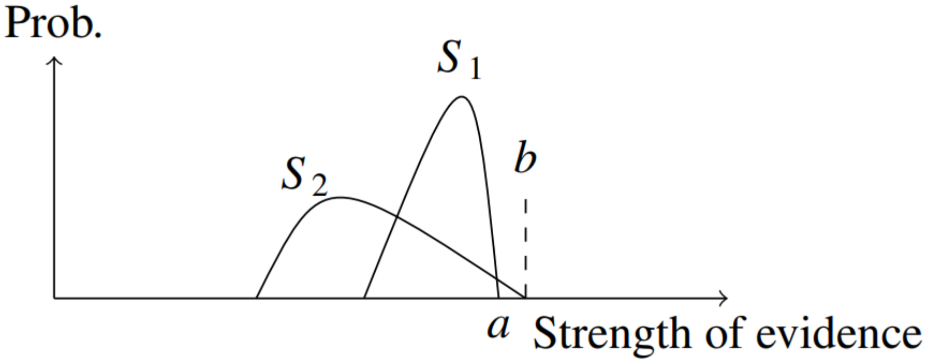

Figure 5. Illustration of the IDEALLY STRONGER interpretation. S_2 is ranked higher than S_1 since the largest available strength of evidence for S_2, labelled *b,* is greater than the largest strength for S_1, labelled *a*.

By "ideally performed" investigations, we obviously mean *really* good investigations methodologically, relative to the given strategy. However, I think that the ideality ought not be understood in an absolute sense but has to be modulated according to what is practically attainable. At least this is true for some variables, the most obvious one being the number of participants (n). Let us assume that n = 250 in a particular investigation judged to be methodologically really good. With n = 260, it would have been even better, and even better still with n = 270. A larger investigation generally provides stronger evidence than a smaller one. But then an ideal investigation would have an immense n. In reality, however, there is a limit on how many subjects can be recruited and handled within given time and cost restraints. Therefore, n cannot be allowed to have an unconstrained influence on the assessment of whether an investigation is (nearly) ideal or not. (Also, there are rare diseases where only a small n can be attained, no matter how much money is available.) If someone says, then, that the ideal S_2 investigation provides stronger evidence than the ideal S_1 investigation, this must be understood as a statement about real or imagined investigations with ns within the bounds of what is practically possible to achieve.

There are also other variables which, if controlled, could raise the quality of an investigation, if ever so slightly, but where such control is unrealistic. For example, we can imagine that in most trials of orally administered drugs, the internal validity would be somewhat raised if all participants were given controlled amounts of food throughout the course of the trial. But normally this is practically impossible. Again, when someone says that the ideal S_2 investigation provides stronger evidence than the ideal S_1 investigation, this has to be understood in relation to what measures are practically possible to take for the strategies in question. If we do not adopt this practically modulated attitude, we will not be able to say that there are any real-world investigations that are (very close to) ideal. On the contrary, all actually performed investigations would be judged far from ideal. This would be a most unhelpful practice, since the upper range of the strength of evidence scale would be rather useless.

It would be possible to make the IDEALLY STRONGER interpretation somewhat weaker while retaining its main idea. Instead of talking about "ideal" (perfect) investigations, one could talk about "good enough" or "well-conducted" investigations, that is, investigations that are close to ideal but not fully ideal. If the ideal S_1 investigation provides stronger evidence than the ideal S_2 investigation, it may be reasonable to assume that the ordering holds also for "very good" or even for simply "good" investigations. But the weaker the condition is made, the less likely it is that what holds for ideal investigations will still hold. This variant can be named thus:

> STRONGER IF GOOD ENOUGH
> An S_1 investigation of sufficient quality provides stronger evidence than any S_2 investigation.

(One might wish for the comparison to be made between investigations of equal quality. From the condition as stated, it follows that an S_1 investigation of sufficient quality provides stronger evidence than an S_2 investigation of equal quality.) An illustration is provided in Figure 6.

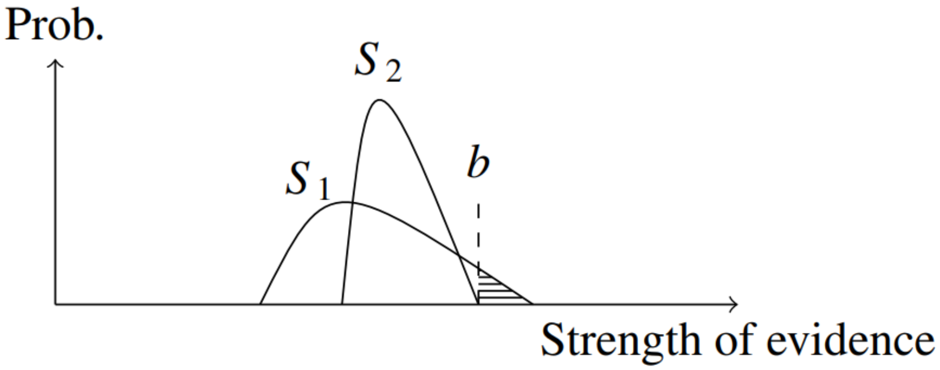

Figure 6. Illustration of the STRONGER IF GOOD ENOUGH interpretation. S_1 is ranked higher than S_2 since there is a portion of S_1 investigations (horizontally shaded) whose strengths of evidence are larger than the maximum strength of evidence for the S_2 strategy, labelled b.

In many cases, the STRONGER IF GOOD ENOUGH interpretation might well be what people have in mind when they discuss the ordering in an evidence hierarchy. For example, two researchers who defended the merits of randomization concluded: "Given the benefits of (well-conducted) randomized trials over (well-conducted) observational studies, where there is any conflict between the results of randomized and observational studies, it seems reasonable to side with the randomized study and assume that its results are more reliable" (Howick and Mebius 2017, 882–883). The idea that one strategy is better than the other when a comparison is made between investigations of the same high quality seems to reflect the STRONGER IF GOOD ENOUGH interpretation.

There is some work that would have to be done before it would be possible to convincingly judge whether one strategy should be above another according to STRONGER IF GOOD ENOUGH ordering. Research quality would have to be operationalized and what counts as "sufficient quality" would have to be established. This is, of course, more or less what some critical appraisal tools try to achieve already; it is just that none seems to depart explicitly from any of the four order relation interpretations delineated here.

It would be possible to find an even weaker interpretation by mixing STRONGER IF GOOD ENOUGH with TYPICALLY STRONGER. This could give, for example: an S_1 investigation of sufficient quality typically provides stronger evidence than any S_2 investigation. The only difference from STRONGER IF GOOD ENOUGH is the addition of the word "typically." Since already the STRONGER IF GOOD ENOUGH interpretation is challenging with respect to how and where to draw the qualitative line, the addition of "typically" makes the task even more complex. However, this interpretation may well have been intended by some evidence hierarchy discussants.

Is the GRADE framework compatible with IDEALLY STRONGER and/or with its weaker relatives? It is not so easy to tell. The fact that the GRADE guidelines allow for both upgrading and downgrading of the assessed strength of evidence ("quality of evidence" in the GRADE terminology) could be used as an argument against saying that GRADE adheres to the IDEALLY STRONGER interpretation. If the default corresponded to ideal investigations, the argument goes, then we would expect only downgrading from this ideal to be allowed.

What about the weaker STRONGER IF GOOD ENOUGH interpretation? Here, we cannot make the identical argument. If the default ranking reflects what is true for investigations that are good enough (but not necessarily perfect), then both upgrading and downgrading would be clearly permissible. However, for STRONGER IF GOOD ENOUGH to be a more plausible interpretation than TYPICALLY STRONGER, one would want some explicit indication from the GRADE rule system and guidelines that points in this direction. I have not found any. In a GRADE paper that explains the rating system, the default rating for an RCT and an observational study, respectively, is simply called "initial quality of a body of evidence" (Balshem et al. 2011, 404, Table 3). There is no hint that the ranking would hold only for investigations of sufficient quality, or that only investigations that are good enough are at all eligible for further processing and assessment in the GRADE algorithms. Also note that an RCT starts out with an initial "high" quality of evidence in GRADE, whereas an observational study starts out with "low" quality of evidence (Guyatt, Oxman, Akl et al. 2011, 386). Evidence from an observational study can be promoted to "high" quality in subsequent steps (passing through an intermediate step called "moderate" quality). The latter fact is somewhat in tension with the idea that the initial rating would reflect the quality of "good enough" observational studies: how could they be good enough if there is so much room for upgrading? I take it that the implicit GRADE reasoning is that RCTs are considered to provide stronger evidence than observational studies due to the defining methodological differences between these two strategies: an RCT is an experiment and features randomization, whereas an observational study is not a proper experiment and does not include randomization. If this is a correct way of understanding GRADE, then the TYPICALLY STRONGER interpretation is the best fit for GRADE, as it implies that RCTs provide stronger evidence than observational studies in a general way, not just restricted to some qualitative top tier of investigations. The STRONGER IF GOOD ENOUGH interpretation cannot be dismissed with absolute certainty, however.

4. Discussion

I offer the four interpretations (with variants) to those who wish to understand what evidence hierarchies mean. For medical evidence hierarchies already published, can we now tell which interpretation is correct? Unfortunately, the answer is no. For GRADE, I have argued that the TYPICALLY STRONGER interpretation seems the most plausible. But there are other systems for evidence assessment, and to the extent that they are using evidence hierarchies, it is mostly unclear which interpretation is the intended one. As time has passed, though, I think there has been a movement away from the NON-OVERLAPPING STRONGER interpretation, particularly in decision contexts where time and considerable resources are available. If it has not already happened, it is now time that we stop taking it as a serious contender in more advanced evidence assessment tasks.

To clarify which interpretation is the intended one seems to be important as soon as the merits or drawbacks of an evidence hierarchy are discussed. It is strange that it has not been done more frequently. It would be useful, first, for the reason already given: to promote clarity of argument, we need to know what we are talking about. Also, I believe that the question of order interpretation could improve related debates. I am thinking primarily of the debate on whether RCTs should be ranked above observational studies. Some discussants have claimed that RCTs and observational studies of the same treatment/disease coupling seem to give practically identical results; therefore, the argument goes, RCTs ought not be placed above observational studies in the evidence hierarchy. The relevance of this argument depends, first, on which RCTs and observational studies have been included in the empirical assessments, and, secondly, on what interpretation of the evidence hierarchy is assumed to be the right one. It may well be the case, for example, that RCTs could occupy a position above observational studies according to the IDEALLY STRONGER interpretation, even though reviews of RCTs and observational studies with wide quality inclusion criteria show essentially no differences in results, for the latter results would not be very relevant according to the IDEALLY STRONGER interpretation.

For transparency, I shall repeat here that methodological assessments and arguments underlie any proposed ordering; for example, the order where RCTs are above observational studies. My point is only that the order, and the arguments for that order, may not be enough for full clarity. The interpretation(s) of the order relation should be made explicit. Any such interpretation should follow from, and be supported by, methodological arguments. But unless the order interpretation is made clear, the risk of discussants talking past one another will be great.

As mentioned early in this article, the overall attempt to clarify what the order in an evidence hierarchy means has two main aspects. First, what does a claim of "stronger" evidence mean, more precisely? Secondly, if the first question has been answered, what are we allowed to do with different pieces of evidence once we know that they can be associated with different strengths in an evidence hierarchy? We have been busy answering the first question, and potential answers are provided by the four interpretations (with variants). The answer to the second question would be particularly interesting if we would like to build models for evidence aggregation. Unfortunately, from answering the first question very little follows that is relevant to the second. For example, from the assertion that a strategy S_1 can be sorted above strategy S_2, according to the IDEALLY STRONGER interpretation, we are not much aided if we face the task of aggregating evidence that emanates from S_1 investigations with evidence from S_2 investigations. The IDEALLY STRONGER condition is just about (practically speaking) ideal investigations, but the investigations from which we have evidence may be very far from ideal (cf. Stegenga 2014, 315). Which one is best: an RCT with certain biases or an observational study with certain (other) biases? This the IDEALLY STRONGER interpretation does not tell.

And, generally, the mere sorting of investigative strategies according to some (explicit or implicit) rule is relatively unhelpful for the evidence aggregation task. It is not surprising, then, that aggregation rules and guidelines have not yet—at least, not to my knowledge—been proposed on the basis of the order properties of evidence hierarchies. The GRADE guidelines contain crude aggregation rules, but these are not based on properties ascribed to the two-item hierarchy from which the whole GRADE procedure starts out. Rather, the aggregation rules are the same irrespective of whether the evidence originates from RCTs or from observational studies (for example, Guyatt, Oxman, Akl et al. 2011, 386).[15]

Stegenga (2014) has expressed strong scepticism about all uses of evidence hierarchies. With respect to the use of evidence hierarchies for facilitating the aggregation of evidence, I essentially share his sceptical conclusion (2014, 315, 318). But Stegenga is not entirely clear about which interpretation he is discussing: some formulations suggest that he is thinking of the STRONGER CP interpretation, others that he is thinking of the IDEALLY STRONGER interpretation. I have shown other options to be available, too. Also, even if evidence hierarchies would rarely (or never) be helpful for aggregation tasks, this does not mean that they are useless in all respects. An evidence hierarchy could be appealed to for assessing the need for more research using particular strategies, for example. But again, which interpretation is used matters for whether such a use is realistic and fruitful.

5. Conclusion

An evidence hierarchy is an ordered list, but the meaning of the order has to be explained, whether the hierarchy is supposed to be used as a rule of thumb, or in a more elaborate evidence assessment procedure. Of the four main interpretations presented, NON-OVERLAPPING STRONGER and STRONGER CP are implausible in the medical context of recommending the best treatments. The TYPICALLY STRONGER and IDEALLY STRONGER interpretations (with variants) are more plausible; indeed, I believe that TYPICALLY STRONGER is the best fit for GRADE. Even if an interpretation were established as the correct one, a lot of additional work would be needed for the evidence hierarchy to be useful for evidence aggregation procedures. However, evidence hierarchies could also be used for other purposes. Proponents and adversaries of evidence hierarchies are equally obliged to specify the order interpretations they are assuming in their arguments. This would further the continuing debate.

[15] A discussion about the reasonableness of GRADE's overall aggregation procedure falls outside the scope of this article. Also, I will not discuss whether it is appropriate for GRADE's evidence hierarchy to consist of only two strategies (RCTs and observational studies) in view of the fact that mechanistic evidence seems to be strong under certain circumstances.

References

Andrews, Jeff, Gordon Guyatt, Andrew D. Oxman, Phil Alderson, Philipp Dahm, Yngve Falck-Ytter, Mona Nasser, Joerg Meerpohl, Piet N. Post, Regina Kunz et al. 2013. "GRADE Guidelines: 14. Going from Evidence to Recommendations: The Significance and Presentation of Recommendations." *Journal of Clinical Epidemiology* 66, no. 7: 719–725. https://doi.org/10.1016/j.jclinepi.2012.03.013.

Andrews, Jeff, Holger J. Schünemann, Andrew D. Oxman, Kevin Pottie, Joerg J. Meerpohl, Pablo Alonso Cello, David Rind, Victor M. Montori, Juan Pablo Brito, Susan Norris et al. 2013. "GRADE Guidelines: 15. Going from Evidence to Recommendation: Determinants of a Recommendation's Direction and Strength." *Journal of Clinical Epidemiology* 66, no. 7: 726–735. https://doi.org/10.1016/j.jclinepi.2013.02.003.

Balshem, Howard, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris et al. 2011. "GRADE Guidelines: 3. Rating the Quality of Evidence." *Journal of Clinical Epidemiology* 64, no. 4: 401–406. https://doi.org/10.1016/j.jclinepi.2010.07.015.

Bluhm, Robyn. 2009. "Some Observations on 'Observational' Research." *Perspectives in Biology and Medicine* 52, no. 2: 252–263. https://doi.org/10.1353/pbm.0.0076.

Broadbent, Alex. 2019. *Philosophy of Medicine*. Oxford: Oxford University Press.

Brunetti, Massimo, Ian Shemilt, Silvia Pregno, Luke Vale, Andrew D. Oxman, Joanne Lord, Jane Sisk, Francis Ruiz, Suzanne Hill, Gordon H. Guyatt et al. 2013. "GRADE Guidelines: 10. Considering Resource Use and Rating the Quality of Economic Evidence." *Journal of Clinical Epidemiology* 66, no. 2: 140–150. https://doi.org/10.1016/j.jclinepi.2012.04.012.

Concato, John. 2004. "Observational *versus* Experimental Studies: What's the Evidence for a Hierarchy?" *NeuroRx* 1, no. 3: 341–347. https://doi.org/10.1602/neurorx.1.3.341.

Evidence-Based Medicine Working Group. 1992. "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine." *Journal of the American Medical Association* 268, no. 17: 2420–2425. https://doi.org/10.1001/jama.1992.03490170092032.

Grossman, Jason and Fiona J. MacKenzie. 2005. "The Randomized Controlled Trial: Gold Standard, or Merely Standard?" *Perspectives in Biology and Medicine* 48, no. 4: 516–534. https://doi.org/10.1353/pbm.2005.0092.

Gugiu, P. Cristian and Mihaiela Ristei Gugiu. 2010. "A Critical Appraisal of Standard Guidelines for Grading Levels of Evidence." *Evaluation & the Health Professions* 33, no. 3: 233–255. https://doi.org/10.1177%2F0163278710373980.

Guyatt, Gordon, Andrew D. Oxman, Elie A. Akl, Regina Kunz, Gunn Vist, Jan Brozek, Susan Norris, Yngve Falck-Ytter, Paul Glasziou, Hans deBeer et al. 2011. "GRADE Guidelines: 1. Introduction: GRADE Evidence Profiles and Summary of Findings Tables." *Journal of Clinical Epidemiology* 64, no. 4: 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, David Atkins, Jan Brozek, Gunn Vist, Philip Alderson, Paul Glasziou, Yngve Falck-Ytter and Holger J. Schünemann. 2011a. "GRADE Guidelines: 2. Framing the Question and Deciding on Important Outcomes." *Journal of Clinical Epidemiology* 64, no. 4: 394–400. https://doi.org/10.1016/j.jclinepi.2010.09.012.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, David Rind, PJ Devereaux, Victor M. Montori, Bo Freyschuss, Gunn Vist et al. 2011b. "GRADE Guidelines: 6.

Rating the Quality of Evidence—Imprecision." *Journal of Clinical Epidemiology* 64, no. 12: 1283–1293. https://doi.org/10.1016/j.jclinepi.2011.01.012.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, James Woodcock, Jan Brozek, Mark Helfand, Pablo Alonso-Coello, Paul Glasziou, Roman Jaeschke, Elie A. Akl et al. 2011c. "GRADE Guidelines: 7. Rating the Quality of Evidence—Inconsistency." *Journal of Clinical Epidemiology* 64, no. 12: 1294–1302. https://doi.org/10.1016/j.jclinepi.2011.03.017.

Guyatt, Gordon H., Andrew D. Oxman, Regina Kunz, James Woodcock, Jan Brozek, Mark Helfand, Pablo Alonso-Coello, Yngve Falck-Ytter, Roman Jaeschke, Gunn Vist et al. 2011d. "GRADE Guidelines: 8. Rating the Quality of Evidence—Indirectness." *Journal of Clinical Epidemiology* 64, no. 12: 1303–1310. https://doi.org/10.1016/j.jclinepi.2011.04.014.

Guyatt, Gordon H., Andrew D. Oxman, Victor Montori, Gunn Vist, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, Ben Djulbegovic, David Atkins, Yngve Falck-Ytter et al. 2011. "GRADE Guidelines: 5. Rating the Quality of Evidence—Publication Bias." *Journal of Clinical Epidemiology* 64, no. 12: 1277–1282. https://doi.org/10.1016/j.jclinepi.2011.01.011.

Guyatt, Gordon H., Andrew D. Oxman, Nancy Santesso, Mark Helfand, Gunn Vist, Regina Kunz, Jan Brozek, Susan Norris, Joerg Meerpohl, Ben Djulbegovic et al. 2013. "GRADE Guidelines: 12. Preparing Summary of Findings Tables—Binary Outcomes." *Journal of Clinical Epidemiology* 66, no. 2: 158–172. https://doi.org/10.1016/j.jclinepi.2012.01.012.

Guyatt, Gordon H., Andrew D. Oxman, Shahnaz Sultan, Jan Brozek, Paul Glasziou, Pablo Alonso-Coello, David Atkins, Regina Kunz, Victor Montori, Roman Jaeschke et al. 2013. "GRADE Guidelines: 11. Making an Overall Rating of Confidence in Effect Estimates for a Single Outcome and for All Outcomes." *Journal of Clinical Epidemiology* 66, no. 2: 151–157. https://doi.org/10.1016/j.jclinepi.2012.01.006.

Guyatt, Gordon H., Andrew D. Oxman, Shahnaz Sultan, Paul Glasziou, Elie A. Akl, Pablo Alonso-Coello, David Atkins, Regina Kunz, Jan Brozek, Victor Montori et al. 2011. "GRADE Guidelines: 9. Rating up the Quality of Evidence." *Journal of Clinical Epidemiology* 64, no. 12: 1311–1316. https://doi.org/10.1016/j.jclinepi.2011.06.004.

Guyatt, Gordon H., Andrew D. Oxman, Gunn Vist, Regina Kunz, Jan Brozek, Pablo Alonso-Coello, Victor Montori, Elie A. Akl, Ben Djulbegovic, Yngve Falck-Ytter et al. 2011. "GRADE Guidelines: 4. Rating the Quality of Evidence—Study Limitations (Risk of Bias)." *Journal of Clinical Epidemiology* 64, no. 4: 407–415. https://doi.org/10.1016/j.jclinepi.2010.07.017.

Guyatt, Gordon H., Andrew D. Oxman, Gunn E. Vist, Regina Kunz, Yngve Falck-Ytter, and Holger J. Schünemann. 2008. "Rating Quality of Evidence and Strength of Recommendations: GRADE: What Is 'Quality of Evidence' and Why Is It Important to Clinicians?" *British Medical Journal* 336, no. 7651: 995–998. https://dx.doi.org/10.1136%2Fbmj.39489.470347.AD.

Guyatt, Gordon and Drummond Rennie, eds. 2002. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago: American Medical Association Press.

Guyatt, Gordon, Drummond Rennie, Maureen O. Meade, and Deborah J. Cook, eds. 2015. *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice*. 3rd edition. New York: McGraw-Hill.

Guyatt, Gordon H., Kristian Thorlund, Andrew D. Oxman, Stephen D. Walter, Donald Patrick, Toshi A. Furukawa, Bradley C. Johnston, Paul Karanicolas, Elie A. Akl, Gunn Vist et al. 2013. "GRADE Guidelines: 13. Preparing Summary of Findings Tables and Evidence Profiles—Continuous Outcomes." *Journal of Clinical Epidemiology* 66, no. 2: 173–183. https://doi.org/10.1016/j.jclinepi.2012.08.001.

Hindenburg, Carl Friedrich. 1796. *Sammlung combinatorisch-analytischer Abhandlungen. Erste Sammlung*. Leipzig: Gerhard Fleischer.

———. 1800. *Sammlung combinatorisch-analytischer Abhandlungen. Zweite Sammlung*. Leipzig: Gerhard Fleischer.

Howick, Jeremy. 2011. *The Philosophy of Evidence-Based Medicine*. Chichester: Wiley Blackwell & BMJ Books.

Howick, Jeremy and Alexander Mebius. 2017. "Randomized Trials and Observational Studies: The Current Philosophical Controversy." In *Handbook of the Philosophy of Medicine*, edited by Thomas Schramme and Steven Edwards, 873–886. Dordrecht: Springer Netherlands.

Jadad, Alejandro R., R. Andrew Moore, Dawn Carroll, Crispin Jenkinson, D. John M. Reynolds, David J. Gavaghan, and Henry J. McQuay. 1996. "Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary?" *Controlled Clinical Trials* 17, no. 1: 1–12. https://doi.org/10.1016/0197-2456(95)00134-4.

La Caze, Adam. 2008. "Evidence-Based Medicine Can't Be…." *Social Epistemology* 22, no. 4: 353–370. https://doi.org/10.1080/02691720802559438.

Nordenstrom, Jorgen. 2007. *Evidence-Based Medicine in Sherlock Holmes' Footsteps*. Malden, MA: Blackwell Publishing.

OCEBM Levels of Evidence Working Group. 2011. "The Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence." Available at https://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf.

Osimani, Barbara. 2014. "Hunting Side Effects and Explaining Them: Should We Reverse Evidence Hierarchies Upside Down?" *Topoi* 33: 295–312. https://doi.org/10.1007/s11245-013-9194-7.

Osimani, Barbara and Fiorenzo Mignini. 2015. "Causal Assessment of Pharmaceutical Treatments: Why Standards of Evidence Should Not Be the Same for Benefits and Harms?" *Drug Safety* 38: 1–11. https://doi.org/10.1007/s40264-014-0249-5.

Stegenga, Jacob. 2014. "Down with the Hierarchies." *Topoi* 33: 313–322. https://doi.org/10.1007/s11245-013-9189-4.

Straus, Sharon E., W. Scott Richardson, Paul Glasziou, and R. Brian Haynes. 2005. *Evidence-Based Medicine: How to Practice and Teach EBM*. 3rd edition. Edinburgh: Elsevier.

Vandenbroucke, Jan P. 2008a. "Observational Research, Randomised Trials, and Two Views of Medical Science." *PLoS Medicine* 5, no. 3: e67, 339–343. https://doi.org/10.1371/journal.pmed.0050067.

———. 2008b. "Observational Research, Randomised Trials, and Two Views of Medical Science." Longer, more detailed version of Vandenbroucke (2008a), available for download at the *PLoS Medicine* website. https://doi.org/10.1371/journal.pmed.0050067.sd001.