Topoi
DOI 10.1007/s11245-013-9189-4

Down with the Hierarchies

Jacob Stegenga

Abstract Evidence hierarchies are widely used to assess evidence in systematic reviews of medical studies. I give several arguments against the use of evidence hierarchies. The problems with evidence hierarchies are numerous, and include methodological shortcomings, philosophical problems, and formal constraints. I argue that medical science should not employ evidence hierarchies, including even the latest and most-sophisticated of such hierarchies.

Keywords Evidence · Causality · Evidence hierarchies · Medicine · Randomized trials · Mechanisms · Amalgamating evidence · Quality assessment tools · RCTs · Meta-analysis

1 Introduction

One of the great challenges of modern medical research is the assessment of causal relations based on very disparate kinds of evidence. Such evidence is often generated from experiments on cells, tissue cultures, and laboratory animals, from mathematical models, epidemiological studies of human populations, controlled clinical trials, and summaries of the primary evidence by formal techniques such as meta-analysis and by social methods such as consensus conferences. Each of these kinds of evidence itself has many variations. Epidemiological studies on humans, for instance, include case–control studies, retrospective cohort studies, and prospective cohort studies.

The so-called evidence-based medicine (EBM) movement organizes and assesses this huge volume and diversity of evidence with 'evidence hierarchies', especially when performing a systematic review of the plethora of available evidence for a given subject. An evidence hierarchy is a rank-ordering of kinds of methods according to the potential for that method to suffer from systematic bias. This rank-ordering is usually determined based on one or very few parameters of study designs. Systematic reviews and specifically meta-analyses are typically held to be at the top of such hierarchies, randomized controlled trials (RCTs) are held to be near the top, non-randomized cohort and case–control studies are held to be lower, and near the bottom are laboratory studies and anecdotal case reports. Table 1 shows a representative example of an evidence hierarchy.

This evidence hierarchy, taken from the Oxford Centre for Evidence-Based Medicine, illustrates the categorical rank-ordering of types of methods commonly used when performing systematic reviews.[1]

The use of hierarchies is widespread in medical science, especially when summarizing evidence from multiple studies in a systematic review. A prominent group in the evidence-based medicine movement emphasizes the employment of evidence hierarchies: "wise use of the

J. Stegenga (✉)
Institute for the History and Philosophy of Science and Technology, University of Toronto, 91 Charles Street West, Toronto, ON M5S 1K7, Canada
e-mail: jacob.stegenga@utoronto.ca
URL: http://individual.utoronto.ca/jstegenga

J. Stegenga
Department of Philosophy, University of Utah, 215 South Central Campus Drive, Carolyn Tanner Irish Humanities Building, 4th Floor, Salt Lake City, UT 84112, USA

[1] I have extracted this table from a slightly more detailed version found on the website of the Oxford Centre for Evidence-Based Medicine (http://www.cebm.net/index.aspx?o=1025, accessed Feb 15, 2013).

Published online: 22 November 2013

Table 1 Evidence hierarchy of the Oxford Centre For Evidence-Based Medicine

Level of Evidence	Method
1a	Systematic review (with homogeneity) of RCTs
1b	Individual RCT (with narrow confidence interval)
1c	All or none
2a	Systematic review (with homogeneity) of cohort studies
2b	Individual cohort study (including low quality RCT)
2c	'Outcomes' Research
3a	Systematic review (with homogeneity) of case–control studies
3b	Individual case–control study
4	Case-series (and poor quality cohort and case–control studies)
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research, or 'first principles'

literature requires a sophisticated hierarchy of evidence" (Karanicolas et al. 2008). Examples of such evidence hierarchies include those of the Oxford Centre for Evidence-Based Medicine, the Scottish Intercollegiate Guidelines Network (SIGN), and The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group.[2] As Howick (2011b) notes, the fundamental view of evidence according to the EBM movement is based on evidence hierarchies.

In this paper I criticize the use of evidence hierarchies. I argue that such hierarchies should not be used to assess evidence when evaluating causal relations in medicine. Several of the problems with evidence hierarchies have been formulated in detail by others.[3] Some recent evidence hierarchies, and philosophical defenses of such hierarchies, have attempted to accommodate some of these previously noted problems. The primary novel contribution in what follows is to argue that even these most recent hierarchies ought to be abandoned, and that the methodological innovations associated with these recent hierarchies in fact amount to an outright abandonment of hierarchies altogether (Sect. 8).

A hierarchy is, formally, a partially ordered set, in which the ordering of the elements in the set is based on a particular property.[4] For evidence hierarchies the elements that are ordered are different kinds of methods, and the property on which the orderings are based is usually taken to be the 'internal validity' of a method relative to hypotheses regarding efficacy of tested medical interventions.[5] Graphical representation of hierarchies, such as Table 1, place some elements above or below other elements based on such orderings. There is no absolute significance of the place that a method holds in a hierarchy: obviously not all tokens of methods from the top of an evidence hierarchy are reliable for inferences regarding efficacy (some RCTs are seriously flawed, as I discuss in Sect. 7), and not all tokens of methods from the bottom of an evidence hierarchy are unreliable for inferences regarding efficacy (for example, some physiological research provides reliable grounds for clinical inferences). I am not aware of anyone who defends such an "absolute" interpretation of evidence hierarchies. Rather, the usual understanding of evidence hierarchies is a weaker "relative" interpretation, which holds that tokens of methods from higher on an evidence hierarchy are more reliable than tokens of methods from lower on the hierarchy. The relative interpretation is nevertheless a strong position in that it holds that the ordering of methods is categorical: that is, it holds that tokens of methods from higher on an evidence hierarchy are *necessarily* more reliable, *ceteris paribus*, than tokens of methods from lower on the hierarchy. It is this relative but categorical interpretation of evidence hierarchies against which I direct most of my arguments.

The relative categorical interpretation of evidence hierarchies is often expressed in methodological guidance for medical researchers, authors of systematic reviews, and physicians studying the clinical research literature. For instance, in an article which purported to provide the best way to distinguish effective medical interventions from those which are ineffective or harmful, the article stated that one should "discard at once all articles on therapy that

[2] To see some evidence hierarchies described and defended, see Wilson et al. (1995), Hadorn et al. (1996) and Atkins et al. (2004).

[3] In part I draw on extant arguments by past critics of evidence hierarchies in medicine, including Bluhm (2005), Upshur (2005), Rawlins (2008), Goldenberg (2009), Borgerson (2009), Solomon (2011) and La Caze (2011). For a specific critique of placing meta-analysis at the top of such hierarchies, see Stegenga (2011), and the assumption that RCTs ought to be necessarily near the top of such hierarchies has been criticized by Worrall (2002) and Cartwright (2007) (among many others).

[4] Etymologically, a hierarchy refers to "rule by priests", in which the hierarch is the top ruling priest.

[5] I am taking internal validity to mean, roughly, freedom from systematic error in the method. This is usually contrasted with external validity, which I take to mean, roughly, the validity of extrapolating results from a test situation to a target situation. For a classic statement of these terms, see Cook and Campbell (1979).

are not about randomized trials" (Department of Clinical Epidemiology and Biostatistics 1981). Similarly, a textbook of methodological guidance claims that "if a study wasn't randomised, we suggest that you stop reading it and go on to the next article in your search" (Straus et al. 2005). These statements of methodological guidance are not necessarily committed to an absolute interpretation of evidence hierarchies, since they are not claiming that all token studies from the top of an evidence hierarchy are all high quality, but they are committed to a relative categorical interpretation of evidence hierarchies, because they are claiming that only randomized trials provide reliable evidence for making inferences about effectiveness of medical interventions.

Some groups have begun to develop evidence hierarchies which appear to dispel with the categorical interpretation. One evidence hierarchy, developed by GRADE, assigns evidence to one of four levels (high, moderate, low, and very low), but these assignments can be changed based on several conditions. In the GRADE scheme, RCTs are automatically categorized as 'high quality' and observational studies are automatically categorized as 'low quality', but the level of an RCT can be decreased by at least one level (to 'moderate') if there are particular methodological problems with the RCT, and the level of an observational study can be increased by up to two levels (to 'high') if the evidence from the observational study is especially salient.[6] It follows that, in principle, GRADE permits an evidence ranking that places token observational studies above token RCTs. Thus one might think that my focus on the categorical interpretation of evidence hierarchies amounts to a straw man argument, since at least one evidence hierarchy appears to be non-categorical. However, there are at least three reasons why this is not so.

First, many evidence hierarchies in use today maintain a straightforward commitment to the categorical interpretation, despite the development of more sophisticated hierarchies such as GRADE (I referred to some of these above, including those of SIGN and CEBM). Second, even the more sophisticated hierarchies are fundamentally based on the categorical interpretation: their starting point for assessing methods is simply the level assigned to a type of method in the hierarchy, and the degree of freedom for modifying this assessment based on relevant methodological properties is limited. Third, the coarse-grained assignment of 'high quality' to a method based on the single property of randomization neglects a vast amount of information that pertains to how compelling that method is, regardless of subsequent modifications to the level assignment. I explain this further in Sect. 8. Moreover, as I also argue in Sect. 8, once one abandons the categorical interpretation of evidence hierarchies, there is compelling reason to think that one has abandoned evidence hierarchies altogether. The considerations that purport to warrant more sophisticated evidence hierarchies in fact warrant abandoning the use of evidence hierarchies.

I begin by granting, for the sake of argument, that for a particular type of hypothesis a principled justification can be provided for a particular hierarchy of evidence types. Even if this is granted, however, I argue that evidence users employ real evidence tokens rather than ideal evidence types, and the warrant for a hierarchy of evidence types may not apply to real evidence tokens (Sect. 2). For different hypothesis types, different evidence hierarchies may be warranted (Sect. 3). As a tool for assessing evidence, hierarchies are crude devices relative to other tools on offer, which include quantitative scales and checklists based on more numerous and more relevant parameters than those included in standard evidence hierarchies (Sect. 4). As methods for amalgamating diverse evidence, hierarchies are non-starters (Sect. 5). The principled reasons proposed for standard evidence hierarchies are inadequate (Sect. 6). Even if we grant that an evidence hierarchy has a well-defined 'top'—a type of evidence deemed categorically superior to others placed lower in the hierarchy— within the type of evidence often said to be at the top (usually from RCTs or meta-analyses), in practice there is wide variability in the quality of evidence (Sect. 7). Some philosophers and methodologists have begun to develop more sophisticated hierarchies in an attempt to accommodate some of these criticisms, but in Sect. 8 I argue that any vestiges of evidence hierarchies that remain should be abandoned. In short, I argue that evidence hierarchies should not be used for assessing causal relations in medicine.

2 Evidence Users Employ Evidence Tokens

Even if principled arguments could warrant an evidence hierarchy, that hierarchy would be constituted by ideal *types* of evidence. Evidence employed by users—policy makers, physicians, patients—is constituted by evidence *tokens*: real instantiations of the ideal types. Principled arguments which warrant a hierarchy of evidence types would not necessarily warrant a hierarchy of evidence tokens, because real tokens of evidence do not necessarily possess the properties of ideal types of evidence which warrant a hierarchy of ideal types of evidence. Ideal RCTs are a great idea, but some real RCTs are dreadful.

The evidence type typically thought to be near the top of evidence hierarchies—evidence from randomized

[6] The evidence from an observational study might be especially salient, for instance, if there were a very strong association between the purported cause and its effect, and there were no obvious threats to the internal validity of the study. See Vandenbroucke (2008).

controlled trials (RCTs)—is powerful, at least in principle. Cartwright (2010) shows that, given a probabilistic theory of causality and some very strong assumptions about the structure of an RCT, if the probability that an effect of interest (E) is achieved in the group in which the purported cause (C) was administered—that is, p(E|C)—is greater than the probability of E in the group in which a placebo or some other non-C factor was administered for comparison—that is, p(E|∼C)—then the evidence from the RCT deductively implies that C causes E. One critical assumption that is necessary for this inference is that the two groups being compared must be homogeneous with respect to all other causes of E not including C.[7] An *ideal* RCT satisfies this strong assumption, because the random allocation of subjects into the group in which the intervention (C) is administered and into the group in which the placebo or non-C factor is administered guarantees that the two groups will be homogeneous with respect to all other causes of E not including C. Other types of evidence employed in biomedical research, such as case–control studies and laboratory research on pathophysiological mechanisms, lack the ability to deductively entail causal conclusions (at least at the clinical level, because laboratory experiments can reveal mechanistic causes). This provides some warrant for placing ideal RCTs at the top of an (ideal) evidence hierarchy.

Real RCTs, on the other hand, do not necessarily satisfy the strong assumptions required to show that the evidence from an RCT deductively entails a causal conclusion. Real RCTs are flawed in various ways, and these flaws can render the assumptions regarding ideal RCTs unwarranted—specifically, randomization does not guarantee that all other causes of E not including C are distributed equally between the experimental group and the control group (Worrall 2002). In Sects. 6 and 7 I survey some of the arguments showing that RCTs do not satisfy the strong ideal assumptions. If the strong ideal assumptions are not warranted, then the evidence from an RCT does not deductively entail a causal conclusion. Like other empirical methods, then, a real RCT merely provides some fallible inductive evidence that C is a cause of E. In short, arguments that warrant a hierarchy of ideal evidence types do not warrant a hierarchy of real evidence tokens.

Moreover, RCTs can be flawed in ways that go beyond threats to internal validity. Such flaws can include a poor operationalization of the outcome of interest, an unrepresentative selection of subjects, and a misleading analysis and presentation of results.

For an example of a poor RCT, consider the RECORD trial, which was intended to provide evidence that rosiglitazone (Avandia) is safe. In the early 2000s some smaller trials had suggested that rosiglitazone causes cardiovascular disease and death. This motivated the manufacturer of rosiglitazone, GlaxoSmithKline, to initiate the RECORD trial. The primary measured outcome in this RCT was a composite outcome: all hospitalizations and deaths from any cardiovascular cause. This gerrymandered outcome very likely included many cardiovascular hospitalizations and deaths which were not caused by rosiglitazone, and thus the observed relative difference in the frequency of cardiovascular hospitalizations and deaths between the group of subjects taking rosiglitazone and the group of subjects taking placebo was minimized. Moreover, the outcome included 'hospitalizations', and since a hospitalization is a complex social and economic event, and the trial occurred in dozens of countries with diverse social and economic settings, the measured results had a higher amount of variability than otherwise would have been the case. Finally, the RECORD trial employed numerous criteria to specify which subjects would be included in and excluded from the RCT. The result of these criteria was that 99 % of the subjects were Caucasian (despite the fact that the trial took place in dozens of countries), and the subjects in the trial were healthier than the equivalent demographic in the sampled population (for example, trial subjects had a heart attack rate of about 40 % of the equivalent demographic in the sampled population).

3 Different Hypothesis Types, Different Hierarchies

Even if principled arguments could warrant an evidence hierarchy for a particular kind of hypothesis, these arguments would not necessarily warrant an evidence hierarchy for other kinds of hypotheses.[8] The typical kind of hypothesis for which the usual evidence hierarchies are said to apply are *this intervention works somewhere* (namely, in controlled experimental settings). Other and arguably more important kinds of hypotheses include *this intervention will work for us*, and *this intervention causes harm*, which will usually require types of evidence not usually thought to be high on standard evidence hierarchies.

[7] These other causes are sometimes called 'confounding causes'. A standard probabilistic definition of causality is: C causes E iff p(E|C and K_i) > p(E| ∼ C and K_i), where K_i are the potential confounding causes. See Cartwright (1979) for a canonical, but slightly different, definition.

[8] In this section and elsewhere I speak of evidence that provides support for various kinds of hypotheses. This way of construing the goal of medical research might appear to be misleading, since often the goal of clinical trials is to determine the strength of a causal relation, as estimated by so-called 'effect sizes'. However, ultimately the goal of medical research is to provide evidence for hypotheses regarding the potential effectiveness of medical interventions. The outcomes of clinical trials, often measured by effect sizes, provide evidence relevant to such hypotheses.

Consider the kind of hypothesis *this intervention will work for us*. For such hypotheses it is not enough to have evidence for the different kind of hypothesis *this intervention works somewhere*. This is a point that Cartwright has been forcefully urging (see, for example, Cartwright 2007). The latter kind of hypothesis might be supported by evidence from an RCT. But when implementing the intervention outside the experimental setting, the causal structure of the target population might be significantly different than the causal structure of the experimental situation (this is the old problem of external validity), and worse, the very implementation of the intervention might modify the causal structure of the target population (this is perhaps especially true when the intervention is one of social policy, as many of Cartwright's examples illustrate, but may also be the case with certain kinds of medical interventions). Different kinds of evidence may be required to overcome this. Thus the evidence hierarchy that may be optimal for *this intervention works somewhere* may not be the same as the evidence hierarchy that is optimal for *this intervention will work for us*.

We ought to be especially concerned with the kind of hypothesis *this intervention causes harm*. This kind of hypothesis will typically require a different evidence hierarchy than the kind of hypothesis *this intervention works somewhere*, because even if principled arguments could warrant an evidence hierarchy with RCTs or meta-analyses of RCTs at the top for the latter kind of hypothesis, the best evidence for many hypotheses of the former kind does not come from RCTs or meta-analyses of RCTs. This is for numerous reasons. There is a direct trade-off between the power of RCTs to detect benefits of medical interventions (providing evidence for the *works somewhere* kind of hypothesis) and the power of RCTs to detect harms of medical interventions (providing evidence for the *causes harm* kind of hypothesis). This trade-off is generated by multiple properties of the design of trials, including the kinds of subjects included or excluded from the trial, the kinds of outcomes measured, and the kinds of analyses performed.[9] The vast majority of RCTs in medical research maximize the power to detect benefit at the expense of the power to detect harm. The majority of serious harms caused by medical interventions are detected by so-called Phase IV post-approval studies, which are almost always limited to observational analyses of anecdotal clinical reports.[10] Thus, for hypotheses of the kind *this intervention causes harm*, medical research is limited to evidence from methods not typically placed near the top of mainstream evidence hierarchies (a compelling case can be made that medical research should overhaul its methods for detecting harmful outcomes of medical interventions—an option that may be appealing but unrealistic in the near future).

4 Hierarchies are Crude Evidence Assessment Tools

Even if principled arguments could warrant an evidence hierarchy, the very nature of a hierarchy constrains quality comparisons between tokens of evidence to ordinal rankings based on relatively few parameters. More sophisticated evidence assessment tools—quality assessment tools, or QATs—permit quality evaluations using numerical scales or checklists which are based on a large number of parameters of medical studies. Most QATs share several elements, including questions about whether or not subjects were randomized to experimental groups in a medical study, whether or not the subjects and experimenters were blinded to the treatment protocol (and if so, how), whether or not there was a sufficient description of subject withdrawal from the study groups, whether or not particular statistical analyses were performed, and whether or not the researchers had financial relationships with corporate sponsors.

Dozens of such QATs are now available. Some are very simple, evaluating only a handful of parameters of medical studies, while others are more complex, evaluating up to 35 parameters of medical studies. For example, (Cho and Bero 1994) developed a QAT with 24 questions, with most of the answers limited to 'yes', 'partial', 'no', or 'not applicable'. These questions are about the design of a study (whether or not it is randomized, placebo-controlled, etc.), whether or not the design of the study was relevant to the research question at hand, whether or not subjects and investigators were blinded to the allocation of the subjects, whether or not the statistical analyses were appropriate, and so on. From these questions an overall numerical score is computed.

Note the relative sophistication of most QATs compared to the simplicity inherent in evidence hierarchies: QATs are able to account for multiple properties of medical studies, and the structure of QATs permits investigation of the relative importance of one property of medical studies compared with another property. Moreover, the evaluation of evidence from medical studies using QATs is usually quantitative. In Sect. 8 I argue that based on such considerations, even the best evidence hierarchies are not as good

[9] See Vandenbroucke (2008) for a discussion of some of these trade-offs, and for a similar defense of the view that different kinds of hypotheses might require different evidence hierarchies.

[10] For more discussion of the view that different hypothesis types require different kinds of studies, see Borgerson (2008).

as typical QATs. The relative sophistication and quantitative nature of QATs should not lead us to overestimate their rigor—elsewhere I describe QATs in more detail and discuss several philosophical implications and problems associated with QATs (Stegenga, forthcoming). Nevertheless, QATs show just how crude evidence hierarchies are, since evidence hierarchies are limited by their very structure to assessing medical studies based on a small number of properties of medical studies (usually randomization), and the assessment itself is limited to an ordinal ranking of kinds of studies.

5 Hierarchies are Poor Methods of Amalgamating Evidence

The evidence hierarchies on offer in modern medical research purport to account for the wide diversity of evidence often available for many hypotheses. Each level of an evidence hierarchy is constituted by a different kind of evidence. These evidence hierarchies are supposed to provide epistemic guidance when faced with diverse evidence. Thus one might think that the evidence hierarchies are methods of amalgamating diverse evidence into an overall measure of confirmation for a hypothesis of interest, akin to other formal methods of amalgamating evidence, such as meta-analysis. They are, however, nothing of the sort.

The guidance that normally comes along with evidence hierarchies is to ignore evidence generated from methods low on the hierarchy, rather than consider it together with evidence generated from methods higher on the hierarchy. Regardless, even if the guidance was to somehow consider the evidence generated from methods low on the hierarchy jointly with evidence generated from methods higher on the hierarchy, the hierarchies themselves lack a substantive technique for doing so. The little formal structure present in evidence hierarchies is too crude to adequately compare or amalgamate evidence of diverse kinds. As discussed above, the evidence hierarchies in the early evidence-based medicine movement categorically ranked quality of evidence by method type, with RCTs at the top and cohort and case–control studies lower. More recent evidence hierarchies, such as that developed by GRADE, allow the assessment of evidential quality to be modified based on additional properties of studies. Nevertheless, even the newer hierarchies do not propose a technique for how the evidence from different kinds of studies should be integrated into an overall measure or confirmation or disconfirmation of a hypothesis.[11]

The way that evidence hierarchies are usually applied is by simply ignoring evidence that is thought to be lower on the hierarchies and considering only evidence from RCTs (or meta-analyses of RCTs). There is a compelling case to be made, however, that evidence from multiple levels of the usual evidence hierarchies should be considered when assessing a causal relation in medicine. The Russo-Williamson thesis, for example, holds that to warrant a causal hypothesis one needs both evidence at a clinical level which supports the causal hypothesis (they call this probabilistic evidence or 'difference-making' evidence) *and* evidence at a mechanistic level which provides an account of how the clinical level evidence came about.[12] When assessing causal relations in medicine we ought to take into account multiple kinds of evidence, but evidence hierarchies possess no compelling method to manage this.

6 Trouble at the Top, in Principle

At the top of all evidence hierarchies in medicine developed thus far are randomized controlled trials (RCTs) or systematic reviews (typically meta-analyses). However, quality among RCTs and meta-analyses is highly variable. A method can hardly be a gold standard if it is highly variable. (Real gold in the namesake gold standard is, after all, extremely homogeneous).

The best justification for placing RCTs at the top of evidence hierarchies is that RCTs minimize certain forms of bias, especially what is known as 'selection bias'. As discussed above, a key assumption required of medical studies in order to deductively warrant a causal conclusion is that the two groups being compared must be homogeneous with respect to all other causes of E not including C. If a medical study is prone to selection bias then the causal homogeneity assumption is very likely not satisfied. Proponents of evidence hierarchies often claim that RCTs guarantee that the causal homogeneity assumption is satisfied. However, as (Worrall 2002) argues, RCTs cannot guarantee this. Randomization does not even render this assumption very probable. Rather, a single iteration of randomization makes it likely that some unspecified subset of confounding causes (the total composition of which is almost always unknown) are more-or-less balanced between the groups, rather than the total set of confounding causes. Thus, one of the most frequently cited justifications for placing RCTs near the top of standard evidence hierarchies is inadequate. Moreover, selection bias is just one

[11] See Douglas (2012) for a discussion of methods for amalgamating diverse kinds of evidence. For an interesting discussion of judgments required in assessing evidence in medicine, see Kelly and Moore (2011).

[12] See Russo and Williamson (2007), Illari (2011), and Leuridan and Weber (2011) for arguments emphasizing the importance of mechanisms in warranting causal hypotheses, and for a critical response see Howick (2011a).

of many kinds of bias pervasive in medical research, and so even if a clinical trial could minimize selection bias, it may still have many other systematic biases.

Meta-analyses—specifically meta-analyses of RCTs—are usually held to be at the top of evidence hierarchies. The intuitive rationale for this is that although one particular medical study might be flawed in various ways or might be too small to detect a significant effect of an intervention, when pooled together the flaws in multiple studies might wash out or the pooled sample size might be large enough to detect an effect of the intervention. However, as I argue in detail elsewhere (Stegenga 2011), meta-analysis suffers from multiple methodological problems which render its results liable to be influenced by idiosyncratic subjective factors. This potential for bias makes meta-analysis malleable—multiple contradictory conclusions can be reached on the same hypothesis by different scientists performing their own respective meta-analyses.

Thus, the methods normally thought to be at the top of evidence hierarchies in medicine—RCTs and meta-analyses—are not as good at minimizing bias and constraining assessments of hypotheses as they are often made out to be.

7 Trouble at the Top, in Practice

The mixed quality of RCTs and meta-analyses have been empirically demonstrated using techniques for assessing evidence more sophisticated than evidence hierarchies (namely, the QATs discussed above). These meta-level studies of medical research have shown a wide disparity in the quality of RCTs. For example, a research group conducted a systematic review of 107 RCTs about a particular medical intervention, using three popular QATs (Hartling et al. 2011). This group found that allocation concealment was unclear in 85 % of these RCTs, and that the vast majority of the RCTs were at high risk of bias.

Another group randomly selected eleven meta-analyses involving 127 RCTs on medical interventions in various health domains (Moher et al. 1998). This group assessed the quality of the 127 RCTs using QATs, and found the overall quality to be low: only 15 % reported the method of randomization, and even fewer showed that subject allocation was concealed. Such examples could be easily multiplied. Some medical scientists have gone so far as to claim that most published research findings are simply false (e.g., Ioannidis 2005, 2008, 2011), and this includes results of studies from the top of standard evidence hierarchies.

The notion of a rigid evidence hierarchy is dubious when token examples from the top of standard evidence hierarchies are so often of low quality. At the very least such examples demonstrate that the absolute interpretation of evidence hierarchies is wrong. Above, though, I argued that few people maintain an absolute interpretation of evidence hierarchies. Do such examples show that the *relative* interpretation of evidence hierarchies is also wrong? This stronger conclusion does not immediately follow. Recall that the relative interpretation holds only that tokens of methods from higher on an evidence hierarchy are more reliable than tokens of methods from lower on the hierarchy. The mere fact that some RCTs and meta-analyses are of low quality does not in itself call into doubt the relative interpretation, since it is at least possible that all tokens of those methods typically thought to be lower in evidence hierarchies (e.g. case–control studies) are even less reliable than the above examples of low quality RCTs. This, however, is extremely implausible. In principle a study could lack the desirable property of randomization but have many other desirable properties, such as large size, proper controls, and subject allocation concealment, and be more compelling than a poor RCT. Indeed, there are many tokens of excellent non-randomized studies. For example, the first evidence that suggested that smoking causes lung cancer came from large non-randomized observational studies.

8 Abandoning Hierarchies

The use of evidence hierarchies in some domains of modern medical research—most notably, when performing systematic reviews—is ubiquitous. I have argued that this is unwarranted. It is revealing to consider the most sustained defense of standard evidence hierarchies employed in evidence-based medicine (that I am aware of), found in Howick (2011b), and to consider the formal implications of the most sophisticated evidence hierarchies in use today. Both amount to the abandonment of evidence hierarchies *simpliciter*.

Howick claims that employing evidence hierarchies is a good strategy in most cases, but he notes that for some medical interventions we have strong confidence in their effectiveness despite a lack of evidence from methods near the top of standard evidence hierarchies (Howick calls this a 'paradox'). The resolution of this paradox, according to Howick, is to consider "all sufficiently high-quality evidence to be weighted together in support of a hypothesis that a treatment caused a patient-relevant outcome." This, Howick suggests, is a minor revision to standard evidence hierarchies which preserves the general commitment to their use.

However, Howick's suggestion is both more revisionary and more problematic than he suggests. Why one should consider only high-quality evidence, and not, say, medium-

quality evidence, in most cases, is not addressed.[13] What constitutes high-quality evidence, and whether or not mechanistic evidence and evidence from non-randomized medical studies is included in the set of high-quality evidence, is precisely what is at issue.[14] How the various kinds of evidence should be 'weighted together' is unstated. Why one ought to consider evidence 'in support of a hypothesis'—without requiring the consideration of evidence against a hypothesis—is presumably an oversight. What distinguishes the standard cases—in which the use of evidence hierarchies is fine—from the paradoxical cases is unclear. Howick's proposal is rife with problems.[15]

Putting these problems aside, one could interpret Howick's suggestions, and any evidence hierarchy consistent with his suggestions (such as that of GRADE), as an outright abandonment of evidence hierarchies. Howick gives conditions for when mechanistic evidence and evidence from non-randomized studies should be considered, and also suggests that sometimes evidence from RCTs should be doubted. If one takes into account methodological nuances of medical research, in the ways that Howick suggests or otherwise, then the metaphor of a hierarchy of evidence and its utility in assessing hypotheses appears at best irrelevant, and at worst misleading.

As I discussed in Sect. 1, some 'evidence hierarchies' such as GRADE employ more than one property to rank methods, and are an example of the approach Howick advocates. Formally, the use of more than one property to rank methods implies that one is not *relying* on a partially ordered set based on a single property, and thus one has abandoned the use of hierarchies as tool for *grounding* the ranking of methods. Of course, trivially, the use of more than one property to rank methods implies that one is still *developing* a partially ordered set based on multiple properties (since one is ranking methods), and thus one is making hierarchies *as a result of* the multi-property ranking of methods. Regardless, this amounts to the abandonment of evidence hierarchies as employed in evidence-based medicine, since the standard evidence-based medicine view has been that the hierarchies provide a *grounding* for rankings of methods.

Moreover, the use of n properties to rank methods is formally equivalent to a scoring system based on n properties which discards any information that exceeds what is required to generate a ranking. Scoring systems, such as the QATs discussed in Sect. 4, generate scores that are measured on scales more informative than ordinal scales (such as interval, ratio, or absolute scales). From any measure on one of these supra-ordinal scales, a ranking can be inferred on an ordinal scale, but not vice versa (from a ranking on an ordinal scale it is impossible to infer measures on supra-ordinal scales).[16] The inference from a supra-ordinal measure to an ordinal measure involves discarding any information beyond mere orderings. Thus the 'quasi-hierarchies' such as GRADE provide evaluations of evidence which are *necessarily less informative* than evaluations provided by QATs.

Another important difference between systems like GRADE and most available QATs is that with GRADE particular studies *begin* with an assignment of quality based on a single property, and that assignment can then subsequently be modified by other properties, whereas with QATs particular studies are *first* evaluated by the multiple properties deemed relevant and only *then* are they assigned a quality rank. In principle the number of properties employed is no different between the two kinds of systems, though in fact GRADE uses a very small number of relevant properties, whereas some QATs employ a large number of relevant properties which provides for a more-detailed evaluation of the evidence.

[13] Howick appeals to a 'principle of total evidence' in defense of his proposal. In fact a principal of total evidence would require one to consider not just high-quality evidence, but all evidence. Leuridan (m.s.) presents a nuanced discussion of the principal of total evidence and the various ways it can be interpreted in the context of medical research.

[14] Mechanistic reasoning is permitted, according to Howick, when the mechanism which is appealed to is 'not incomplete'. But critics of evidence hierarchies do not suggest appealing to evidence from methods purported to be lower on the evidence hierarchy when that evidence is sketchy. Moreover, it is fine to say that all high-quality evidence should be considered when that evidence is concordant, but hard cases are when plausible evidence from methods on various levels of an evidence hierarchy conflict with each other.

[15] See Bluhm (2011) for a thorough critique of Howick (2011b).

[16] This typology of scales is standard (see, for exposition, Suppes and Zinnes (1962)). Consider the following examples of four different kinds of comparisons:

(i) Beth claims that the food at Kiribati Kuisine is better than at Tahitian Treats
(ii) The temperature inside Kiribati Kuisine is 20 °C while the temperature inside Tahitian Treats is 22 °C
(iii) Kiribati Kuisine has been in business 5 years longer than Tahitian Treats
(iv) Kiribati Kuisine has fewer items on its menu than does Tahitian Treats

The scales of these measurements are, respectively, ordinal (i), cardinal (ii), ratio (iii), and absolute (iv). Any positive transformation to a measure of Beth's food tastes will preserve the information in (i). Only a positive linear transformation will preserve the information in (ii)—for instance, if we switched to the Kelvin scale. Similarly, only a positive linear transformation will preserve the information in (iii)—for instance, if we switched to a scale of weeks instead of years—and there is a natural zero point: the date of business inception. Only an identity transformation will preserve the information in (iv): the actual number of items on each menu. From a measure on a ratio scale—take the one in (ii), for example—we can infer a measure on an ordinal scale—in this example, that it is colder inside Kiribati Kuisine than it is inside Tahitian Treats.

Finally, another important difference is based on the weight that can be assigned to individual properties of studies. Because GRADE starts with a quality assignment based on one property and takes other properties into account by subsequent modifications of the quality assignment (shifting the assignment up or down one or two levels), the weights that can be assigned to various properties are limited to unit values proportional to the number of possible quality assignments divided by the number of assessed properties. In GRADE there are four quality levels and (for the sake of simplicity) n properties, and the level switches are limited to one or two per property, and thus each property can count for either $1/n$ or $2/n$ of the final quality level assignment. Thus the weight assigned to each property is highly constrained. With QATs, on the other hand, the weight assigned to each property is completely open, and can be set based on rational arguments regarding the respective importance of the various properties without arbitrary constraints imposed by the structure of the scoring system.

In short, at present some philosophers and methodologists are suggesting a departure from standard evidence hierarchies (relative categorical hierarchies) in favor of more sophisticated hierarchies (relative non-categorical hierarchies). This may be due in part to some of the extant criticisms of standard hierarchies canvassed throughout this paper. I have argued that a non-categorical evidence hierarchy amounts to an abandonment of the very principle that motivated evidence hierarchies in the first place. Regardless, any vestiges of evidence hierarchies that remain should be excised, since these vestiges constitute evidence assessment schemes which are unreasonably constraining and less informative than other schemes now available.

9 Conclusion

I have raised several arguments against the use of evidence hierarchies. Nevertheless, the original motive for the use of evidence hierarchies remains: modern medical research has an enormous volume and diversity of evidence for many hypotheses, the various kinds of evidence have different inductive strengths and weaknesses, and somehow this messy mass of evidence must be analyzed to provide guidance for action. That is what evidence hierarchies are meant to do, and thus without evidence hierarchies some other way of analyzing the volume and diversity of evidence is necessary. Proposals abound. (Bluhm 2005) suggests the notion of an evidence *network* as an alternative metaphor for structuring evidence, akin to Bradford Hill's suggestion that when assessing causal hypotheses one ought to appeal to a plurality of kinds of evidence. Similarly, (Borgerson 2008) notes a renewed interest in methodological pluralism in medical research, and cites other proposals for metaphors meant as alternatives to evidence hierarchies, including an evidence *matrix* (Petticrew and Roberts 2003), and Cartwright (m.s.) proposes an evidence *pyramid*. I briefly noted that QATs are another possibility. These alternatives remain undeveloped, but have the potential to generate better ways to manage the great volume and diversity of evidence in medical research than is offered by evidence hierarchies.

Acknowledgments I am grateful to Phyllis Illari, Federica Russo, and two anonymous reviewers for detailed commentary on earlier drafts. Financial support was provided by the Banting Postdoctoral Fellowships program administered by the Social Sciences and Humanities Research Council of Canada.

References

Atkins D, Best D, Briss PA, Group GW (2004) Grading quality of evidence and strength of recommendations. BMJ 328:1490

Bluhm R (2005) From hierarchy to network: a richer view of evidence for evidence-based medicine. Perspect Biol Med 48(4):535–547. doi:10.1353/pbm.2005.0082

Bluhm R (2011) Jeremy Howick: the philosophy of evidence-based medicine. Theor Med Bioeth 32(6):423–427. doi:10.1007/s11017-011-9196-7

Borgerson K (2008) Valuing and evaluating evidence in medicine. PhD diss

Borgerson K (2009) Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. Perspect Biol Med 52(2):218–233. doi:10.1353/pbm.0.0086

Cartwright N (1979) Causal laws and effective strategies. Nous 13:419–437

Cartwright N (2007) Are RCTs the gold standard? BioSocieties 2(1):11–20. doi:10.1017/s1745855207005029

Cartwright N (2010) What are randomised controlled trials good for? Philos Stud 147:59–70

Cho MK, Bero LA (1994) Instruments for assessing the quality of drug studies published in the medical literature. JAMA 272(2):101–104

Cook TD, Campbell DT (1979) Quasi-experimentation: design and analysis issues for field settings. Houghton Mifflin, Boston

Department of Clinical Epidemiology and Biostatistics, M. U. H. S. C (1981) How to read clinical journals: V: to distinguish useful from useless or even harmful therapy. Can Med Assoc J 124(9):1156–1162

Douglas H (2012) Weighing complex evidence in a democratic society. Kennedy Inst Ethics J 22(2):139–162

Goldenberg MJ (2009) Iconoclast or creed? Objectivism, pragmatism, and the hierarchy of evidence. Perspect Biol Med 52(2):168–187. doi:10.1353/pbm.0.0080

Hadorn DC, Baker D, Hodges JS, Hicks N (1996) Rating the quality of evidence for clinical practice guidelines. J Clin Epidemiol 49:749–754

Hartling L, Bond K, Vandermeer B, Seida J, Dryden DM, Rowe BH (2011) Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. PLoS One 6(2):e17242. doi:10.1371/journal.pone.0017242

Howick J (2011a) Exposing the vanities—and a qualified defense—of mechanistic reasoning in health care decision making. Philos Sci 78(5):926–940

Howick J (2011b) The philosophy of evidence-based medicine. Wiley, Oxford

Illari PM (2011) Mechanistic evidence: disambiguating the Russo–Williamson thesis. Int Stud Philos Sci 25(2):139–157. doi:10.1080/02698595.2011.574856

Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2(8):e124. doi:10.1371/journal.pmed.0020124

Ioannidis JP (2008) Why most discovered true associations are inflated. [Review]. Epidemiology 19(5):640–648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis JP (2011) An epidemic of false claims. Competition and conflicts of interest distort too many medical findings. Sci Am 304(6):16

Karanicolas PJ, Kunz R, Guyatt GH (2008) Point: evidence-based medicine has a sound scientific base. [Editorial]. Chest 133(5):1067–1071. doi:10.1378/chest.08-0068

Kelly MP, Moore TA (2011) The judgement process in evidence-based medicine and health technology assessment. Soc Theory Health 10(1):1–19

La Caze A (2011) The role of basic science in evidence-based medicine. Biol Philos 26(1):81–98. doi:10.1007/s10539-010-9231-5

Leuridan B, Weber E (2011) The IARC and mechanistic evidence. In: Illari PM, Russo F, Williamson J (eds) Causality in the sciences. Oxford University Press, NewYork, pp 91–109

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, M Moher, Klassen TP (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352(9128):609–613. doi:10.1016/s0140-6736(98)01085-x

Petticrew M, Roberts H (2003) Evidence, hierarchies, and typologies: horses for courses. J Epidemiol Community Health 57(7):527–529

Rawlins M (2008) De Testimonio: on the evidence for decisions about the use of therapeutic interventions. Royal College of Physicians, London

Russo F, Williamson J (2007) Interpreting causality in the health sciences. Int Stud Philos Sci 21:157–170

Solomon M (2011) Just a paradigm: evidence-based medicine in epistemological context. Eur J Philos Sci 1(3):451–466. doi:10.1007/s13194-011-0034-6

Stegenga J (2011) Is meta-analysis the platinum standard? Stud Hist Philos Biol Biomed Sci 42:497–507

Stegenga J (forthcoming) Quality of information in clinical research. In: Illari PM, Floridi L (eds) The philosophy of information quality. Springer

Straus SE, Richardson WS, Glasziou PP, Haynes RB (2005) Evidence-based medicine: how to practice and teach, 3rd edn. Elsevier Churchill Livingstone, London

Suppes P, Zinnes JL (1962) Basic measurement theory. Institute for mathematical studies in the social sciences, Technical Report No. 45

Upshur RE (2005) Looking for rules in a world of exceptions: reflections on evidence-based practice. Perspect Biol Med 48(4):477–489. doi:10.1353/pbm.2005.0098

Vandenbroucke JP (2008) Observational research, randomised trials, and two views of medical science. PLoS Med 5(3):e67. doi:10.1371/journal.pmed.0050067

Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt G (1995) Users' guides to the medical literature. VIII. How to use clinical practice guidelines. B. what are the recommendations and will they help you in caring for your patients? The evidence-based medicine working group. JAMA 274(20):1630–1632. doi:10.1001/jama.1995.03530200066040

Worrall J (2002) What evidence in evidence-based medicine? Philos Sci 69:S316–S330