An Evidence-Hierarchical Decision Aid for Ranking in Evidence-Based Medicine

Jürgen Landes

Abstract This chapter addresses the problem of ranking available drugs in guideline development to support clinicians in their work. Based on a pragmatic approach to the notion of evidence and a hierarchical view on different kinds of evidence this chapter introduces a decision aid, HiDAD, which draws on the multi criteria decision making literature. Properties, modifications and applicability of HiDAD are discussed.

1 Introduction

Evidence in medicine has been a hot topic ever since the late 1980's. Special interest has been paid to the questions "What constitutes best or legitimate evidence in medicine?", "How does one amalgamate evidence for medical decision making?" and "How important is the most important kind of evidence?", see, e.g., (Clarke et al., 2014; GRADE Working Group, 2004; Howick and co workers, 2011; Osimani, 2014b,a; Russo and Williamson, 2007; Sackett et al., 2000; Worrall, 2007a). This chapter builds on previous work providing answers to these questions and introduces a ranking heuristic for comparing drugs termed *HIerarchical Decision AiD* (HiDAD) and addresses, to various degrees, all these three questions. As an illustrating example I shall consider the following decision problem: a medical body is re-writing its guidelines to treat migraines. The problem arises to rank drugs for treating migraines to guide clinicians in their work.

By bringing formal machinery, which was developed in decision science, to bear on this ranking problem I aim to create a decision heuristic which is transparent to the stakeholders (patients, doctors, guideline developers, politicians, drug developing and manufacturing companies, journalists etc.) and to direct attention of

Jürgen Landes
Munich Center for Mathematical Philosophy, LMU Munich, Ludwigstr. 31, D-80539 Munich, Germany e-mail: Juergen.Landes@lrz.uni-muenchen.de

the evidence-in-medicine movement to parts of the, so far, under-appreciated decision science literature. I aim to achieve these aims by delineating the importance of different kinds of evidence and by making all subjective judgements during the decision making process explicit and transparent.

Clearly, to successfully complete this ranking task all the *available evidence* ought to be taken into account and, ceteris paribus, the more evidence that is available the better the final decision. Philosophically, these principles have been expressed as the *Principle of Total Evidence* (Carnap, 1947) and as the *Value of Knowledge Theorem*, see (Savage, 1954, Chapter 7) and (Skyrms, 1990, Chapter 4). Unsurprisingly, hiding of important information can have disastrous consequences, e.g., in the infamous Vioxx case (Horton, 2004; Jüni et al., 2004; Krumholz et al., 2007; McGauran et al., 2010). The ranking problem is further complicated by reporting biases in medicine distorting the available evidence, see (Bes-Rastrollo et al., 2013; Every-Palmer and Howick, 2014; McGauran et al., 2010), which are not addressed here.

While the "available" part in "available evidence" is uncontentious, there is much philosophical debate about the notion of "evidence", e.g., (Kelly, 2015; Reiss, 2015; Williamson, 2015). Here, I am concerned with a concrete decision problem and take a pragmatic approach. I shall take all information to be evidence which *can on its own or jointly with other information conceivably influence the decision problem* of ranking drugs. This notion of evidence is hence depending on the decision problem at hand and the decision making entity's epistemic state at this time.[1] This approach allows, even compels, one to take into account and aggregate all possibly relevant information; in particular evidence comes in different kinds or sorts such as Randomised Controlled Trials (RCTs), cohort studies, expert testimony, case reports, etc.[2]

There are further reasons arising from the concrete decision problem to take such a liberal view on what constitutes evidence: Ideally, studies would license the same inferences for the studied population and the target population. In reality, studies are not conducted on the *entire population of interest* but on a much smaller number (even too small a number) of patients, see (Chan and Altman, 2005, p. 1160), see (Button et al., 2013) for this problem in neuro-science, (Doll and Peto, 1980) in cancer research, (Etz and Vandekerckhove, 2016, p. 10) in psychology and a philosophical discussion of this problem in (Worrall, 2007a, p. 992) and (Bertamini and Munafó, 2012) for a general discussion. Additionally, studied populations, in particular, RCTs often fail to be representative for the target population due to strict patient inclusion criteria, see (Revicki and Frank, 1999) and (Upshur, 1995, p. 483).

[1] Information known to be false or irrelevant is thus ignored. Which information is deemed relevant and which is deemed irrelevant is a complicated question outside the scope of this contribution. The answers will depend on the epistemic state, as well as cognitive limitations and the exact framing of the decision problem.

[2] In the more applied sciences, the term *information fusion* rather than *evidence amalgamation* or *evidence aggregation* is often used. Definitions of the term *information fusion* are surveyed in (Boström et al., 2007). Further often-used terms are "research synthesis" and "evidence synthesis", see also Section 2.1.3.

Furthermore, observational and/or cohort studies are sometimes much larger than the largest conducted RCT.

Yet another concrete reason for taking such a liberal view is that smaller studies using different protocols rather than a large single study are a good way to counter biases due to subject selection, study design, and execution strategy in a single study, as argued by epidemiologists in (Borm et al., 2009). Philosophically, this idea has been expressed as the *Variety-of-Evidence Thesis* which states that the more varied the body of evidence is, ceteris paribus, the more informative it is. Earman called this a "truism of methodology" (Earman, 1992, p. 77).

With such a liberal view on evidence and a highly diverse body of evidence, there is a lot of evidence to be taken into account. One ubiquitous intuition regarding different kinds of evidence is that some kinds or sorts of evidence matter more than others. This intuition manifests itself in an ever-growing number of evidence hierarchies which rank kinds of evidence from most to least important, see (Canadian Task Force on the Periodic Health Examination, 1979; Evans, 2003; GRADE Working Group, 2004; Howick and co workers, 2011) and see http://cjblunt.com/hierarchies-evidence/ for a much more complete list of close to 100 evidence hierarchies. A kind of evidence is termed "level" in (Howick and co workers, 2011).

I develop the decision aid based on this construal of evidence, a hierarchical (ranked) view on kinds of evidence and a qualitative approach to appreciate evidence I develop the decision aid HiDAD.

The rest of this chapter is organised as follows: Next, I spell out the decision problem I tackle in detail and argue that current approaches are inadequate for resolving it. Then, I show HiDAD can be used to help tackle this decision problem, discuss some of its properties, possible variations and its applicability. Finally, I conclude.

2 The Decision Problem

The formal decision problem I here consider is a regulatory body tasked with writing guideline recommendations.[3] This regulatory body aims to:

> Rank a number of available drugs to treat migraines in non-pregnant adults according to their respective outcome given all the available evidence.

The term *outcome* is to be widely understood, it includes treatment efficacy, side effects (negative as well as positive effects) and monetary as well as non-monetary

[3] These recommendations are intended to *guide* doctors in their daily work. I emphatically do not want to suggest that a recommendation of a regulatory body ought to be followed at all times. There are good reasons to deviate from general medical guidelines when it comes to the treatment of individual patients. Patients have individual circumstances such as: co-morbidities, known or suspected (drug)-intolerances and treatment preferences as well as outcome preferences. For deciding on a treatment in an individual patient at a particular time, these patient-specific circumstances ought to matter, too.

costs.[4] I denote the set of available drugs to be ranked by $\mathscr{A} = \{A_1,\dots,A_a\}$. The option of not treating migraines at all is, of course, also a possible course of action. For ease of exposition, I shall assume that this *null act* is included in the set of alternatives $\mathscr{A}$. The literature on decision making also knows the words "options", "acts" or "actions" instead of "alternative".

A ranking is simply some formal way of expressing that some alternatives are judged (based on the available evidence) to be more preferable than some other alternatives. A ranking in this sense is not necessarily transitive, acyclic nor complete, e.g., A_1 might be ranked higher than A_2 which is ranked higher than A_3 which is ranked higher than A_1 (cycle of length three) and A_1 may not be ranked relative to A_4 (ranking is incomplete).

As is tradition, the decision making entity (in this the case the regulatory body) is referred to as the *Decision Maker* (DM). In reality, the DM consists of a number of different individuals with possibly conflicting preferences. I here ignore this layer of complexity and point the reader to the group decision making literature in general and to (Urfalino, 2012) for an analysis of the actual procedures by which group decisions are made by regulatory bodies. How regulatory bodies do approach such a ranking problem is nicely described in (Kelly and Moore, 2012, p. 4-5).

2.1 Related Methodological Work

Closely related work to this approach is the literature on evidence hierarchies, GRADE in particular, and on medical decision support. This literature is discussed next.

2.1.1 Evidence Hierarchies

I now briefly explain why evidence hierarchies alone do not solve the drug ranking problem. Typically, the first level of evidence hierarchies is made up of systematic reviews (many of which, but not all, are meta-analyses). The lower levels are, typically, populated by the case reports and the expert opinions. (Howick, 2011, p. 5) describes a very simple hierarchy with three levels: the first level is made up of randomised trials, the second level is made up of observational studies while the third and lowest level is made up of expert judgement and mechanistic reasoning.

There are two ways of interpreting and subsequently applying current evidence hierarchies.[5] 1) A trumping interpretation of an evidence hierarchy entails that high level evidence *trumps* lower level evidence. That is, the right decision is determined

[4] There is no principled reason for which I could not construe the decision problem as a multi-outcome problem. For migraines, these outcomes might be: hours with headache, headache severity, days of sick leave and adverse events. In order to keep the complexity of the problem and of the presentation manageable, I abstain from doing so.

[5] The GRADE approach is discussed in more detail in Section 2.1.2.

only by high level evidence, if such evidence is available and favors one alternative over another. Lower level evidence hence plays no role in case high level evidence is available and favors one alternative over another. It matters not how strongly the high level evidence favors one alternative over the other, nor does it matter how strongly the reversal is at the lower levels.

2) According to the less strict interpretation, these hierarchies only rank kinds of evidence. This interpretation of evidence hierarchies only entails a *ranking of kinds of evidence* without trumping. That is, a DM faced with a concrete decision problem is only provided with a ranking which assigns every item of evidence a level. In general, these rankings alone will be much too little to help a DM deciding between different alternatives A_k and A_i in $\mathscr{A}$. What the DM is lacking is a clear decision process (guideline) which allows the aggregation of items of evidence falling into different ranking levels.

Concerning 1) Recently, philosophers have argued for putting more weight on the lower levels of evidence hierarchies. Hence, higher level evidence should not automatically trump lower level evidence. In (Clarke et al., 2013, 2014), Clarke et al. argue that the evidence hierarchies currently on the market under-value mechanistic evidence. Worrall questions the uniquely privileged epistemic role of RCTs in (Worrall, 2002, 2007a,b, 2010). Cartwright points to a low external validity of RCTs (Cartwright, 2007; Cartwright and Munro, 2010). Osimani and Vandenbroucke argue in (Osimani, 2014b) and (Vandenbroucke, 2008) that current evidence hierarchies are inadequate to capture potential risks of health interventions and advocate to put more weight on the lower levels of the hierarchies. Stegenga goes as far as calling for an end of all evidence hierarchies in medicine in (Stegenga, 2014). The legal scholar Twinning also expressed his dislikes for hierarchies at (Twinning, 2011, p. 76).

Solomon puts forward the idea that “ranking of evidence is done by reference to the actual, rather than the theoretically expected, reliability of results” (Solomon, 2011, p. 463-464). For current purposes, this means that the hierarchy of evidence to be used depends on the decision problem and the available evidence. Solomon also questions the reliability of using actual RCTs in clinical decision making while acknowledging that this is controversial, (Solomon, 2011, Section 3.2).[6] La Caze argues in (La Caze, 2009) that an evidence hierarchy is best understood as a hierarchy of comparative internal validity.

It is not only philosophers who have criticised the trumping interpretation of evidence hierarchies. Epidemiologists and physicians also worry about hierarchies and evidence amalgamation. Borm et al. have claimed that the best way to evaluate the performance of a treatment is to use multiple, possibly smaller, trials (Borm et al., 2009, p. 711) which can be seen as further arguments against trumping. Upshur has expressed the worry that RCTs may be under-powered for secondary outcome measures and are hence severely limited in informing us about the harm/benefit ratio, see (Upshur, 1995).

[6] Recently, it was alleged that it is impossible to amalgamate of evidence of different kinds (Stegenga, 2013); an appropriate response was provided in (Lehtinen, 2013).

Onakpoya et al. give a long table of approved drugs which were world-wide withdrawn from the market without(!) any high-level evidence (Onakpoya et al., 2016, Table 1). These withdrawals are difficult to square with the trumping interpretation in which low-level evidence is taken into account only if the higher-level evidence does not clearly indicate the drug causes an adverse drug reaction. Under the implicit assumption that it was the right decision to withdraw these drugs, their work speaks against the trumping interpretation.

In sum, I think that the cumulative force of these arguments is strong enough to demonstrate that the strict interpretation of evidence hierarchies is inadequate for this drug ranking problem.

2.1.2 GRADE

The most closely related decision method which grades evidence and subsequently makes recommendations in health care is the GRADE system (GRADE Working Group, 2004; Guyatt et al., 2013, 2008, 2011). Like this approach, GRADE is a bottom-up approach which first draws comparisons between two health interventions based on studies or meta analysis and then aggregates these comparisons to arrive at a recommendation of one health intervention over the other.[7] I agree with Guyatt et al. that the

> [...] merit of GRADE is not that it necessarily ensures reproducible judgments (observers will inevitably differ in close-call situations when rating up or down for individual domains or for the overall confidence per outcome) but that it achieves explicit and transparent judgment. (Guyatt et al., 2013, p. 155)

However, I think that GRADE suffers from a number of drawbacks which make it inadequate for solving the ranking problem. While I would argue that some of these drawbacks are almost insurmountable, I will only aim to establish the significantly weaker claim that taken together these drawbacks make GRADE inadequate for current ranking purposes.

I) GRADE comes with a pre-described evidence hierarchy which only has three levels where evidence best suited to detect adverse drug reactions of newly released drugs (case reports, expert opinions), see (Onakpoya et al., 2016), is entrenched as the least important kind of evidence. The DM hence cannot use her own favourite hierarchy. For example, DMs worried about safety of a newly released drug cannot move case reports and experts opinions up in the hierarchy. They will hence have a hard time to make safety signals weigh heavily.

II) In GRADE, neither upgrading nor downgrading of evidence is clearly operationalised. For example, it is not clear in which cases inconsistency in the data is important, when data is sparse, when directness is major nor when the study quality

[7] Without going into details here, GRADE and HiDAD use similar language to refer to different concepts and techniques.

is (seriously) limited.[8] There is no clear question formulated which upon answering would allow the DM to decide whether to upgrade or downgrade items of evidence. That is, no intuitive scales are given to which a DM may calibrate her judgements.

III) GRADE does not make explicit – with or without further judgement – how the grading of evidence leads to a decision. For example, it is left open whether three studies downgraded for conflict of interest are as valuable as two studies with no conflict of interest, ceteris paribus. That is, it is not clear how the choice of up-grading or down-grading an item of evidence effects the decision making. In other words, GRADE does not offer a way to aggregate these (upgrade/downgrade) judgements to arrive at a resolution of the decision problem. This lack of clarity becomes the more severe the more the DM needs to aggregate heterogeneous judgements. The very recent 'Evidence to Decision Frameworks', see (Alonso-Coello et al., 2016), offer great practical value in how to approach the decision problem in practise. However, they offer no way to resolve the problems discussed here.

IV) GRADE requires an explicit judgement for every study. For a comparison of two health interventions for which there exist a large number of studies, the number of choice points will be large. Given the large number of subjective choices feeding into the overall recommendation, the recommendation made appears to be based more on judgement than on evidence. This seems to be less than ideal in *evidence* based medicine which aims to reduce the number of subjective judgements.[9]

2.1.3 Medical Decision Support Literature for Evidence Amalgamation - A broader Perspective

There exist a number of decision support systems for medical decision making see, e.g., (van Valkenhoef et al., 2013). A readable, though by no means complete overview, can be found in (van Valkenhoef et al., 2013, pp. 463-464). These systems support the aggregation of medical data of one kind; normally RCTs only. The outputs of these systems often are (translatable into) preferences over health care interventions; based on evidence of a single kind. One particular strand of such support systems are rapid reviews which are tools for faster amalgamation of medical evidence, see (Khangura et al., 2014). Practical issues arising from the need for amalgamation are discussed in (Thomas et al., 2013).

A review of medical decision models developed in the UK during a seven years span (1997 - 2003) clearly expresses the need for the amalgamation of evidence of different kinds (Cooper et al., 2005, p. 249)

> Currently, the formal synthesis of evidence tends to be limited to RCT data and applied using standard meta-analysis techniques,[16] where appropriate. However, with a move towards identifying all relevant sources of evidence for model inputs, the application of generalized

[8] The ever-present difficulties from passing from a continuum to a discretisation (of judgements) are another layer of complexity (Guyatt et al., 2013, p. 154-155), which apply equally to GRADE and to HiDAD.

[9] Under the construal of evidence offered here, expert clinical judgement is evidence, too.

> evidence synthesis methods that combine both randomized and non-randomized data are needed. Some methods for generalized evidence synthesis have been proposed,[17-19]...

All three models referred to in this quote are Bayesian models: they use precise numbers to represent the DM's epistemic uncertainty; see also the very recent (Landes et al., 2017). In the present setting, I want to avoid such precise quantification, as I shall explain in Section 2.2 below.

I take the apparent lack of proposals assigning cardinal weights to levels of evidence hierarchies to be evidence that decision makers are not able and comfortable with precise quantifications in this context.

Important episodes in the history of how the amalgamation of evidence (in medicine) became a field on its own is summarised in (Chalmers et al., 2002, p. 25). Chalmers et al. identify DMs as a driving force behind this development

> Consumers of research have begun to point out more forcibly that "atomized", unsynthesized products of the research enterprise are of little help to people who wish to use research to inform their decisions.

Kelly & Moore provide philosophical underpinnings for the amalgamation of evidence

> The idea is that the greater the number of observations, the greater the degree of accuracy about that which is being observed and the greater the chance of the elimination of uncertainty. Further, if multiple studies with multiple results are pooled, then there is an even better chance of the results being averaged out in an optimally accurate way. This is a way of dealing with uncertainty. It recognises explicitly that single observations may be unreliable and that multiple observations offer protection against outliers. (Kelly and Moore, 2012, p. 7)

2.2 *Multi Criteria Decision Making for this Ranking Problem*

I develop HiDAD with the goal in mind to help the DM decide on how to compare two alternatives A_k and A_i given the available evidence. Comparing the evidence for the overall outcome of A_k to that of A_i can be a daunting task, in particular, when, say, the RCTs speak in favor of A_k and the cohort studies and observational studies are more favourable towards A_i while case reports are only available for alternative A_k. In applications, such comparisons of alternatives with such heterogeneous alternatives are not straightforward.

Less difficult are comparisons between two alternatives when attention is *restricted* to one specified kind of evidence. I take it, that in the current ranking problem to be the case that the DM is *able and comfortable* to assess whether one kind of evidence supports the conclusion that a) alternative A_k is much better than an alternative A_i, b) an alternative A_k is better but not much better than A_i c) alternative A_k is as good as A_i or d) that the strength of the evidence of this kind which supports A_k over A_i is *incomparable* to the strength of the evidence of this kind which supports A_i over A_k. This last case, d), might be deemed to be the case when there are a great number of contradictory cohort studies available investigating A_k but no cohort studies have been conducted for alternative A_i.

I call such a comparison restricted to evidence of one particular type a *marginal* comparison.[10] The decision science literature refers to a decision problem which can be analysed in terms of different evaluation criteria as a multi-criteria decision problem. In this ranking problem, the different marginal evidential support relations play the role of evaluation criteria. The reader who is unacquainted with Multi Criteria Decision Making (MCDM) is referred to the excellent overviews (Belton and Stewart, 2002; Bouyssou et al., 2006; Figueira et al., 2005; Keeney and Raiffa, 1993).[11]

Naturally, the interesting cases are the cases in which two alternatives A_i and A_k are judged differently in different criteria. Say, A_i is preferable to A_k in three criteria and A_k is preferred to A_i when evaluated on four other criteria. Not making any recommendations in the face of such heterogeneous evidence is often not an option for a regulatory body. So, a decision making process has to be followed which leads to recommendations for treating migraines.

2.2.1 Landscapes of Multi Criteria Decision Making

One way to divide the multi-criterial decision methodology landscape is along the following fault line. A) Every alternative is first evaluated in every evaluation criterion, these evaluations are then aggregated into one final overall score for an alternative. The overall score of an alternative A_i is then compared to the overall score of other alternatives A_k in order to recommend (or come to) a decision [top-down]. B) The second type of approach is to first draw all marginal comparisons between alternatives and then aggregate these marginal comparisons between alternatives to support the decision making process [bottom-up].

In general, it makes a difference whether a top-down or a bottom-up approach is pursued. For example, a top-down approach assigning every alternative a score which is a real number can easily determine a transitive, acyclic and complete ranking by ranking alternatives according to their score. A bottom-up approach on the other hand aggregates marginal comparisons to determine a ranking between pairs of alternatives. It is highly unclear how to aggregate the marginal comparisons into a ranking over alternatives which is transitive, acyclic and complete in general.

Another way to carve up the set of multi-criterial decision methods is by distinguishing 1) quantitative (cardinal) approaches from those 2) which refrain from

[10] There is no suggestion here that even such limited comparisons are always feasible. I would like to refer the reader to Footnote 14 for further discussion. To help determine marginal comparisons the DM may choose to avail herself to further (medical) decision aids. For example, a) to assess (systematic reviews of) RCTs the DM may use decision support systems put forward in the medical decision literature which were discussed in Section 2.1.3, b) means to make sense of multiple, possibly conflicting, expert opinions are put forward in the literature on judgement aggregation.

[11] The term multi criteria decision *analysis* is also often found in the literature which is, at times, used interchangeably.

using precise quantitative evaluations. Qualitative Decision Theory (QDT) is a different paradigm than the "usual" expected utility maximisation, the latter is due to Savage (Savage, 1954). QDT supports decision making processes by non-cardinal approaches. The term "qualitative decision theory" dates back at least two decades to (Boutilier, 1994; Tan and Pearl, 1994), refer to (Doyle and Thomason, 1999) for a very readable albeit slightly outdated overview and to (Dubois et al., 2002) for a compact contrasting of QDT and expected utility maximisation. A recent overview of qualitative decision rules under uncertainty may be found in (Dubois et al., 2009).[12] Normally, it is not advisable to use a qualitative approach, if quantification is sensible.

2.2.2 Purely Ordinal MCDM

A naïve qualitative decision theoretic approach to the decision problem represents the marginal comparisons of alternatives in terms of "better supported", "equally supported" or "less supported". That is, all marginal comparisons are purely ordinal, it only matters which alternative is better supported by a sub-body of evidence but how much greater the support is, is irrelevant. The arising decision problem is said to be a purely ordinal multi-criterial decision problem.

Dubois et al. showed in (Dubois et al., 2002, Corollary 2) that, given very natural axioms formalising the essence of purely ordinal decision making, purely ordinal multi-criterial decision problems are *oligarchical*. That is, there have to exist oligarchies of criteria which decide the overall comparison of two alternatives. Only if all stronger oligarchies are indifferent between two alternatives, may criteria of lower importance influence the overall ranking of two alternatives.

As discussed in Section 2.1.1, such oligarchical decision making (trumping) is not appropriate for the decision problem under consideration. Rejecting one of the natural axioms of ordinal decision making in an ordinal decision making problem is also not an option I entertain here.

It seems that I am caught between a rock and a hard-place: On the one side I face a complex decision problem with heterogeneous evidence which only supports relatively weak comparative claims and on the other hand the results by Dubois et

[12] A reluctance to use precise numbers has not only manifested itself in the analysis of decision problems but also in the related, but by no means equivalent, epistemological problem of determining rational degrees of beliefs. This reluctance has given rise (among others) to the framework of imprecise probabilities, see (Troffaes and de Cooman, 2014) for a very recent treatment, Dempster-Shafer Theory, see (Shafer, 1976), and fuzzy logic as championed by Dubois and Prade, see (Dubois et al., 1997). In (Shafer and Srivastava, 1990, p. 129), Shafer & Srivastava argued in favor of qualitative approaches [those with "fewer inputs" in their terminology] thusly:

> *When fewer inputs are required, we have a better chance of finding reasonably solid evidence on which to base these inputs, and thus, we have a better chance of producing an overall argument based on evidence rather than mere fancy.*

al. show that there are no decision rules appropriate for purely ordinal multi-criterial decision making.

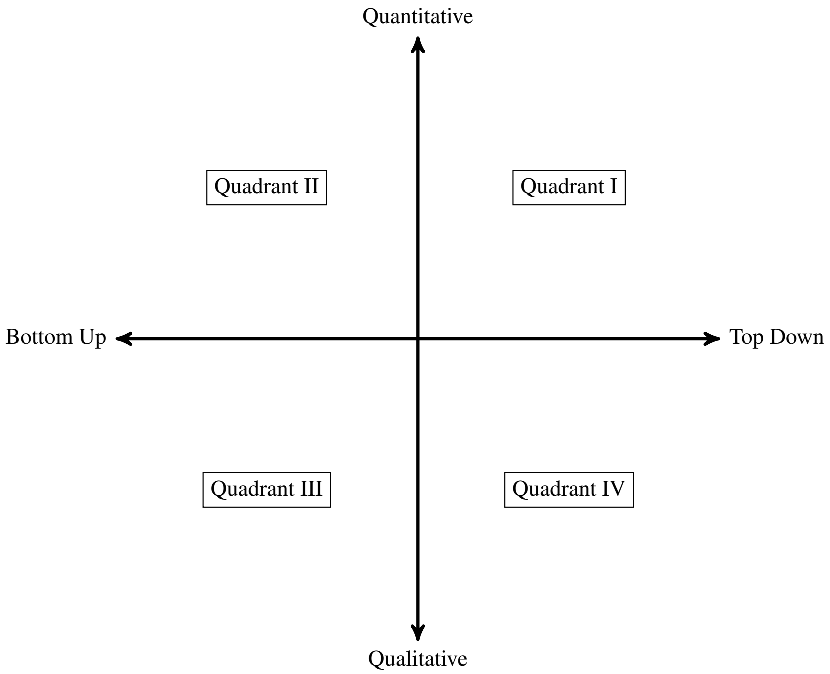

Fig. 1 Coordinate system for multi-criteria decision making.

2.3 *Methodology of this Approach*

I take it as my starting point here, that given the available evidence for the ranking problem, the DM is not able and comfortable to give an overall assessment of an alternative nor is the DM able to articulate meaningful quantitative judgements.[13] Thus, HiDAD first draws comparisons and then aggregates marginal comparisons and it refrains from using sharp quantitative evaluations. It is the decision problem at hand which motivates HiDAD's location in the third quadrant (bottom-up and qualitative) of the coordinate system drawn in Figure 1.

[13] An ideal rational agent, the protagonist of many a philosophical piece, may be in a position to give meaningful precise quantitative assessments. A (group of) human decision makers is in a significantly different epistemic situation. The applicability of HiDAD depending on the DM's situation is discussed in Section 6. Section 6.1 focuses on applications of HiDAD to other problems, while Section 6.2 provides conditions under which HiDAD should not be applied.

I escape the predicaments of purely ordinal multi-criterial decision making by allowing for a final ranking that may be cyclic, incomplete and intransitive and by representing marginal comparisons in a more nuanced, i.e., not purely ordinal, way, see Section 3.1 for an operationalisation of the marginal comparisons.

3 HiDAD

As mentioned above, a number of evidence hierarchies have been put forward which disagree on the kinds of evidence to take into account and how to order kinds of evidence. HiDAD supposes that the DM has already established a hierarchical ordering of kinds of evidence, whatever this ordering may be as long as the ordering is strict and total, i.e., for every two different kinds of evidence one of them is deemed strictly more important than the other. The choice of a particular hierarchy is left to the DM. There is some genuine subjective choice on the DM's part here, e.g., whether case reports and expert clinical judgement combine for a level of the hierarchy, or if case reports and expert clinical judgement constitute levels on their own; in the latter case they also need to be ranked.[14] I opt for the second interpretation of the evidence hierarchy, see Section 2.1.1.

Having set-up the decision problem and spelled out the assumptions I now present HiDAD. To do so I shall suggest solutions to two questions: 1) How can one operationalise the marginal comparisons, i.e., how to elicit the marginal comparisons from the DM? 2) How can one aggregate marginal comparisons into an overall ranking?

The answer to the first question is found by appreciating the available evidence (Section 3.1), the second question is answered in Section 3.2 and Section 3.3. A schematic representation of HiDAD is given in Figure 2.

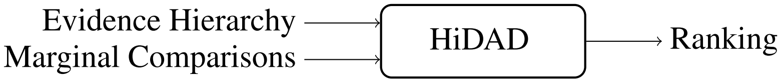

Fig. 2 Given an evidence hierarchy and all marginal comparisons (the input) HiDAD deterministically outputs the overall ranking.

The need for a systematic way of aggregating marginal comparisons arises from the motivation to design a decision aid which requires relatively few subjective judgements as inputs. Surely, there is no absolute need for an systematic aggregation in general and aggregation of marginal comparisons via case-by-case reasoning is possible. Such a method would lack the systematic approach presented here.

[14] Clearly, it may not always be the case that the DM is able and comfortable to do so. This does not mean that HiDAD is wrong, it simply means that it should not be applied in such a case. Mutatis mutandis, the same is true for further assumptions I make: If the assumptions I make do not hold in another concrete decision problem, then HiDAD should not be applied, see also Footnote 10.

Furthermore, if aggregation is performed via case-by-case inferences, then subjective judgements used to determine aggregates are neither explicit nor transparent to stakeholders.

3.1 Evidence Appreciation

The way evidence is appreciated is given in Table 1 and Table 2 using the following bit of notation. Fixing the evidence hierarchy provided by the DM, I denote the sub-body of evidence consisting of all evidence of Level L by $\mathscr{E}_L$. Furthermore, define the sub-body of evidence of the most important L levels by $\mathbb{E}_L := \bigcup_{l=1}^{L} \mathscr{E}_l$. By definition, it holds that $\mathscr{E}_1 = \mathbb{E}_1$. The entire body of evidence is denoted by $\mathbb{E}$.

Marginal Comparison	The sub-body of evidence of the first (most-important) level, $\mathscr{E}_1$, supports the conclusion that when A_i is compared to A_k that
$A_k \gg_1 A_i$	A_k is so much better than A_i that the DM prefers A_k over A_i no matter all the other evidence.
$A_k >_1 A_i$	A_k is better than A_i but not that much better that $A_k \gg_1 A_i$ holds.
$A_k \sim_1 A_i$	A_k and A_i are roughly equally good.
$A_k \ll_1 A_i$	A_i is so much better than A_k that the DM prefers A_i over A_k no matter all other evidence.
$A_k <_1 A_i$	A_i is better than A_k but not that much better that $A_k \ll_1 A_i$ holds.
$A_k \bowtie_1 A_i$	A_k and are A_i incomparable.

Table 1 Operationalisation of the marginal comparisons on the first, most important, level of evidence.

Marginal Comparison	The sub-body of evidence $\mathscr{E}_L$ ($L > 1$) supports the conclusion that when A_i is compared to A_k that
$A_k \gg_L A_i$	$\mathbb{E}_{L-1}$ equally supports A_k and A_i, A_k is so much better than A_i that the DM prefers A_k over A_i no matter all other evidence in $\mathbb{E} \setminus \mathbb{E}_L$.
$A_k >_L A_i$	A_k is better than A_i but not that much better that $A_k \gg_L A_i$ holds.
$A_k \sim_L A_i$	A_k are and A_i are roughly equally good.
$A_k \ll_L A_i$	$\mathbb{E}_{L-1}$ equally supports A_k and A_i, A_i is so much better than A_k that the DM prefers A_i over A_k no matter all other evidence in $\mathbb{E} \setminus \mathbb{E}_L$.
$A_k <_L A_i$	A_i is better than A_k but not that much better that $A_i \gg_L A_k$ holds.
$A_k \bowtie_L A_i$	A_k and are A_i incomparable.

Table 2 Operationalisation of the marginal comparisons on all but the first level of evidence, $L > 1$.

I take it that for every level L and every pair of alternatives A_i and A_k the DM is able and comfortable to judge that A_k and A_i stand in exactly one of the relations $\gg_L, >_L, \sim_L, \ll_L, <_L, \bowtie_L$, that is, I take it that these six relations are exhaustive and mutually exclusive.

Every item of evidence is assigned a level L in the DM's evidence hierarchy. Ceteris paribus, the higher the level, the more influential the item of evidence will be for the decision. Furthermore, different studies on the same level may carry different evidential weights. For example, the external and internal validity as well as the effect size of the studies populating level L feed into the subjective judgement as to which of these six relations holds on Level L. So, while items of evidence only wield influence within their assigned level, they can do so to different degrees. For example, a set of RCTs may be all but ignored due to suspected biases, $A_1 \sim_2 A_2$, while a large group of observational studies on Level 3 may sway the DM to put $A_1 \gg_3 A_2$.

A DM worried about adverse drug reactions may set $A_i \gg_1 A_k$, only in case there are safety RCTs which clearly establish that A_i is safer than A_k. See (Price et al., 2014) for a recent description of design and analysis of safety trials.

Example 1. Consider a case with a large number of high-quality meta-analyses which all show a that A_1's treatment outcomes are significantly better than those of A_2. The DM may hence judge that $A_1 \gg_1 A_2$. HiDAD will hence recommend A_1 over A_2.

In a case in which the available expert opinions (Level 6) regarding A_3 vary from "widely effective and no side effects" over "as good as the standard treatment" to "ineffective with significant adverse drug reactions", while there are no expert opinions on the newly approved drug A_4 available, the DM may judge that $A_3 \bowtie_6 A_4$.

In case there is no evidence on Level L concerning A_k and also no evidence concerning A_i available, a reasonable *DM* will set $A_k \sim_L A_i$.

Regarding the first case in this example, HiDAD does not maintain that higher-level evidence always trumps lower-level evidence, but higher level evidence determines preferences in certain cases. For example, a large number of high-quality meta-analyses (internally and externally valid) which all show a clearly better outcome for A_1 than for A_2. In such a case, HiDAD maintains that A_1 is preferable to A_2.

There are cases in which there are no RCTs available where good observational studies alone are sufficient to swing the pendulum in one way; for a forceful argument of this point see the discussion of the ECMO case in (Worrall, 2007b, Section 2). Hence, it makes sense to formalise decisive evidence on the lower levels, say, $\gg_5$ in HiDAD.

Incomparability occurs when the evidence on Level L regarding two alternatives is strongly heterogeneous and the DM cannot say which alternatives is better supported or whether both alternatives are equally-well supported, see (Aumann, 1962) for an influential piece on the notion of incomparability. In such a case, it seems plausible to me, that higher level evidence is required to determine the DM's preferences. Hence, the lower-level evidence will be of no use, since the evidence which matters most is so heterogeneous and the lower-level evidence does not hold sufficient sway.

A DM has a hard time deciding between two alternatives, if she has equal preference for these two alternatives or if the alternatives are incomparable. In the first case, an assessment of the body of available evidence suggests that both alternatives are equally good; in the second case the available body of evidence does not allow such an assessment. So, while both cases lead to a difficult choice problem, they do so for different epistemic reasons. When further but less influential evidence becomes available, then the epistemic state in both cases may, in general, change in different ways; see further Table 5.

3.2 Aggregation of Comparisons

With the evidence appreciated I now turn to evidence amalgamation of different kinds by aggregating the marginal comparisons. I aggregate the marginal comparisons step-by-step and begin with the first, most important, level. Example 2 illustrates the notation for and the aggregation of comparisons.

I shall use relations $\succ_L, \succeq_L, \approx_L, \prec_L, \preceq_L, \perp_L$ to formalise aggregated comparisons on the first L levels combined. The intended meanings are given in Table 3.

Ranking up to and including level L	The sub-body of evidence $\mathbb{E}_L$ supports the conclusion that when A_i is compared to A_k that
$A_k \succ_L A_i$	A_k is so much better than A_i that the DM prefers A_k over A_i no matter all other evidence in $\mathbb{E} \setminus \mathbb{E}_L$.
$A_k \succeq_L A_i$	A_k is better than A_i but not that much better that $A_k \succ_L A_i$ holds.
$A_k \approx_L A_i$	A_k are and A_i are roughly equally good.
$A_k \prec_L A_i$	A_i is so much better than A_k that the DM prefers A_i over A_k no matter all other evidence in $\mathbb{E} \setminus \mathbb{E}_L$.
$A_k \preceq_L A_i$	A_i is better than A_k but not that much better that $A_k \prec_L A_k$ holds.
$A_k \perp_L A_i$	A_k and are A_i incomparable.

Table 3 Meanings of the relations representing relative evidential support on the first L levels.

With these meanings in place, it is straight-forward to define the first step of the aggregation. Given the marginal comparisons on Level 1 I put forward definitions of $\succ_1, \succeq_1, \approx_1, \prec_1, \preceq_1, \perp_1$ in Table 4.

Next, I inductively define the aggregated comparison relations up to Level L given two ingredients: i) the marginal comparisons on Level L and ii) the overall aggregated comparisons up to and including Level $L-1$ in Table 5.

I now briefly discuss the entries of Table 5.

By the definition of $A_k \succ_{L-1} A_i$ (conclusive evidence for preferring A_k over A_i on the most important $L-1$ levels, $\mathbb{E}_{L-1}$) and $A_k \prec_{L-1} A_i$ (conclusive evidence for preferring A_i over A_k), the marginal comparisons on Level L do not have any influence on the aggregates, that is, $A_k \succ_L A_i$ (respectively $A_k \prec_L A_i$) always holds – no matter the evidence on Level L [second and fifth column of Table 5]. This is in line with the definition of $\succ_{L-1}$ and $\prec_{L-1}$ in Table 3.

Marginal Comparison	Overall Ranking up to Level 1
$A_k \gg_1 A_i$	$A_k \succ_1 A_i$
$A_k >_1 A_i$	$A_k \succeq_1 A_i$
$A_k \sim_1 A_i$	$A_k \approx_1 A_i$
$A_k \ll_1 A_i$	$A_k \prec_1 A_i$
$A_k <_1 A_i$	$A_k \preceq_1 A_i$
$A_k \bowtie_1 A_i$	$A_k \perp_1 A_i$

Table 4 Evidential support relations given marginal comparisons on the first and most important level.

	$A_k \succ_{L-1} A_i$	$A_k \succeq_{L-1} A_i$	$A_k \approx_{L-1} A_i$	$A_k \prec_{L-1} A_i$	$A_k \preceq_{L-1} A_i$	$A_k \perp_{L-1} A_i$
$A_k \gg_L A_i$	$A_k \succ_L A_i$	$A_k \succ_L A_i$	$A_k \succ_L A_i$	$A_k \prec_L A_i$	$A_k \approx_L A_i$	$A_k \perp_L A_i$
$A_k >_L A_i$	$A_k \succ_L A_i$	$A_k \succ_L A_i$	$A_k \succeq_L A_i$	$A_k \prec_L A_i$	*	$A_k \perp_L A_i$
$A_k \sim_L A_i$	$A_k \succ_L A_i$	$A_k \succeq_L A_i$	$A_k \approx_L A_i$	$A_k \prec_L A_i$	$A_k \preceq_L A_i$	$A_k \perp_L A_i$
$A_k \ll_L A_i$	$A_k \succ_L A_i$	$A_k \approx_L A_i$	$A_k \prec_L A_i$	$A_k \prec_L A_i$	$A_k \prec_L A_i$	$A_k \perp_L A_i$
$A_k <_L A_i$	$A_k \succ_L A_i$	*	$A_k \preceq_L A_i$	$A_k \prec_L A_i$	$A_k \prec_L A_i$	$A_k \perp_L A_i$
$A_k \bowtie_L A_i$	$A_k \succ_L A_i$	$A_k \succ_L A_i$	$A_k \perp_L A_i$	$A_k \prec_L A_i$	$A_k \prec_L A_i$	$A_k \perp_L A_i$

Table 5 Overall evidential support relations given marginal comparisons on the Level L and aggregated marginal comparisons up to and including Level $L-1$. The fields marked by * are explained in detail in the text.

For alternatives which are equally supported by the sub-body of evidence $\mathbb{E}_{L-1}$ the marginal comparisons on Level L determine the aggregate preference [fourth column of Table 5].

If the sub-body of evidence $\mathbb{E}_{L-1}$ is such that the strength of the evidence which supports A_k over A_i cannot be compared to the strength of evidence which supports A_i over A_k, then no evidence of less importance can break this incomparability [last column of Table 5].

I now address the third column of Table 5; and by symmetry also the second to last column. If $\mathbb{E}_{L-1}$ supports A_k over A_i but not as much such that $A_k \succ_{L-1} A_i$ holds, then the further evidence on Level L which supports A_k over A_i is enough to tip the balance in favor of A_k over A_i for good. In case the evidence on Level L equally supports A_k and A_i, then the overall ranking of A_k and A_i will remain unchanged. On the other hand, strong evidence in favor of A_i over A_k ($A_k \ll_L A_i$) will lead to a modification of the overall ranking of A_k and A_i; I here suggest to change it to $A_k \approx_L A_i$.

Bottom entry of the third column: A sub-body of evidence $\mathscr{E}_L$ which does not allow a comparison between the evidence in $\mathscr{E}_L$ which supports A_k over A_i and the evidence in $\mathscr{E}_L$ which supports A_i over A_k renders marginal comparisons on lower levels meaningless. It is hence at this Level L that I will have to set the overall ranking of A_k and A_i. Given that $\mathbb{E}_{L-1}$ supports A_k over A_i it appears sensible to set $A_k \succ_L A_i$.

The intuition behind this aggregation is the same as above, in order to break ties between incomparable alternatives higher level evidence is required. In this case, this means that the lower level evidence is – sadly – ignored and the higher level evidence carries the day.

*: In case there is weak evidence in favor of A_i over A_k ($A_k <_L A_i$), then this should not automatically result in a change of the overall ranking since the evidence on Level L is less important than the evidence on higher levels.

I suggest the following procedure to determine *: Let $1 \leq j \leq L-1$ be the level at which the overall ranking switches for the last time in favor of A_k, formally let $1 \leq j \leq L-1$ be maximal such that $A_k \succeq_j A_i$ holds but $A_k \succeq_{j-1} A_i$ fails to hold. If for all $j+1 \leq r \leq L-1$ $A_k \sim_r A_i$ holds, then there is not enough evidence to change the overall ranking and hence $A_k \succeq_L A_i$ holds. But if there exists one $j+1 \leq r \leq L-1$ such that $A_k <_r A_i$ holds and if for all other $j+1 \leq s \leq j-1$ either $A_k \sim_s A_i$ or $A_k <_s A_i$ hold, then $A_k \approx_L A_i$ holds as there are two levels of evidence which support A_i over A_k.[15] For an example see the example below for A_5 and A_6.

Example 2. Consider an evidence hierarchy with five levels. Comparing two alternative drugs $A_1, A_2 \in \mathscr{A}$ such that $A_1 >_1 A_2$ and $A_1 >_2 A_2$ one finds $A_1 \succeq_1 A_2$, $A_1 \succ_2 A_2$ and thus $A_1 \succ_5 A_2$. That is, based on the entire body of evidence, HiDAD ranks A_1 higher than A_2, no matter the evidence on the three lower levels. This may reflect the DM's subjective judgement that, say, RCTs (Level 1) and cohort studies (Level 2) combined, are deemed much more important than case reports (Level 3), expert testimony (Level 4) and animal data (Level 5) combined.

For alternatives A_3, A_4 and the same hierarchy with $A_3 >_1 A_4$, $A_3 \ll_2 A_4$, $A_3 <_3 A_4$, $A_3 <_4 A_4$ and $A_3 <_5 A_4$ one has $A_3 \succeq_1 A_4$, $A_3 \approx_2 A_4$, $A_3 \preceq_3 A_4$, $A_3 \prec_4 A_4$ and finally $A_3 \prec_5 A_4$. Hence, HiDAD ranks A_4 higher than A_3. This is an example of non-trumping: the evidence on the first level is not strong enough to suppress the evidence on the lower levels which suggests that A_4 is preferable to A_3. Lower level evidence has changed the DM's preferences.

Concerning *: Assume that $A_5 >_1 A_6$ and thus $A_5 \succeq_1 A_6$ and assume furthermore that $A_5 <_2 A_6$. Since the evidence on the second level matters less than the evidence on the first level, it seems right to set $A_5 \succeq_2 A_6$. In case there is further evidence supporting A_6 over A_5 in the form of $A_5 <_3 A_6$, then my proposal is to set $A_5 \approx_3 A_6$. That is, the lower level evidence has changed the ranking from a moderately more supported, $\succeq_1$, to equally supported, $\approx_3$. HiDAD's recommendation will, in this example, depend on the lower level evidence.

If and only if either $A_k \preceq_{L-1} A_i$, $A_k \succeq_{L-1} A_i$ or $A_k \approx_{L-1} A_i$ hold, then $A_k \gg_L A_i$, $A_k \ll_L A_i$ and $A_k \bowtie_L A_i$ will lead to a change of the aggregated comparisons. So, strong evidence on lower levels plays an important role in HiDAD.

[15] Note that all other marginal comparisons lead to a change of the overall ranking of A_k and A_i and hence all possible cases have been considered here.

3.3 The Ranking

Having determined evidential support relations on the entire body of evidence, $\mathbb{E}$, I can now determine the overall ranking of alternatives. That is, I now say which recommendations the body of evidence $\mathbb{E}$ supports according to HiDAD. The final ranking relations are given in Table 6. I use $\prec, \approx, \succ, \perp$ to denote these mutually exclusive and exhaustive relations.

Overall Support	The entire body of evidence $\mathbb{E}$ supports the conclusion that
$A_k \succ_L A_i$	A_k is better than A_i, $A_k \succ A_i$.
$A_k \succeq_L A_i$	A_k is better than A_i, $A_k \succ A_i$.
$A_k \approx_L A_i$	A_k are A_i equally good, $A_k \approx A_i$.
$A_k \prec_L A_i$	A_i is better than A_k, $A_k \prec A_i$.
$A_k \preceq_L A_i$	A_i is better than A_k, $A_k \prec A_i$.
$A_k \perp_L A_i$	The strength of the evidence in $\mathbb{E}$ supporting A_k over A_i is incomparable to the strength of evidence in $\mathbb{E}$ which supports A_i over A_k. A_k and A_i are are incomparable, $A_k \perp A_i$.

Table 6 The overall ranking of pairs of alternatives based on the aggregation of all marginal comparisons.

Based on the final ranking the DM then makes the guideline recommendation to doctors treating migraines in non-pregnant adults:

> Use the *maximal elements* of the final ranking.

A maximal element is an alternative A_k such that there is no alternative $A_i \in \mathscr{A}$ with $A_k \prec A_i$. The rationale for this recommendation is simple: The maximal elements appear to be best, given all the available evidence, see Figure 3 for a visualisation.

In case there are no maximal elements, a number of plausibly sensible courses for actions suggest themselves. I shall here only list some these options without endorsing a particular one: i) Exploit information encoded in $\succeq_L$, ii) recommend all alternatives A_k for which the number of other alternatives A_i which are ranked lower ($A_k \succ A_i$) is maximal or iii) recommend all but the minimal elements in the final ranking. An alternative A_k is not minimal in the final ranking, if and only if there exists an alternative A_i which is ranked lower than A_k, $A_k \succ A_i$.

Example 3. Continuing Example 2: HiDAD recommends A_1 over A_2 and A_4 over A_3. Given the ranking depicted in Figure 3 HiDAD recommends the two alternatives A_4 and A_5.

4 Properties of HiDAD

I now briefly discuss properties of the final ranking depending on the marginal comparisons.

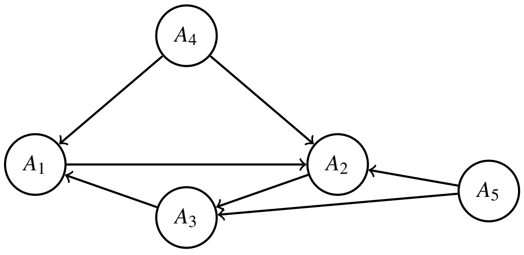

Fig. 3 Example of a final ranking, directed arrows are intended to mean that the target of the arrow is ranked lower than alternative from which the arrow originates. In the depicted configuration, A_4 and A_5 are the only maximal elements (no arrows end there). There are no minimal alternatives, an arrow originates from every alternative. This ranking is cyclic (A_1,A_2,A_3 form a cycle) and intransitive ($A_4 \succ A_2$ and $A_2 \succ A_3$ both hold, but $A_4 \succ A_3$ does not hold).

4.1 Compatibility

Let $A_1, A_2, A_3 \in \mathscr{A}$ be pairwise different alternatives and consider the set of marginal comparisons of A_1 to A_2 and the marginal comparisons of A_1 to A_3. If for all these marginal comparisons A_1 does better when compared to A_3 than when it is compared to A_2, then overall ranking of A_1 relative to A_3 ought to be better or equal than the ranking of A_1 relative to A_2. I now show that this does indeed hold.

To simplify notation I introduce the concept of a better or equal marginal comparison. For a fixed level L, I say that the set of better or equal comparison(s) of

- $\gg_L$ is $\{\gg_L\}$,
- $>_L$ is $\{\gg_L, >_L\}$,
- $\sim_L$ is $\{\gg_L, >_L, \sim_L\}$,
- $<_L$ is $\{\gg_L, >_L, \sim_L, <_L\}$,
- $\ll_L$ is $\{\gg_L, >_L, \sim_L, <_L, \ll_L\}$.

If for all levels L the marginal comparison of A_1 relative to A_3 is better or equal than the comparison of A_1 relative to A_2, I then say that the ordered pair (A_1, A_3) is *marginally better or equal* than the ordered pair (A_1, A_2).

Proposition 1 (Compatibility). *If for all levels L neither $A_1 \bowtie_L A_2$ nor $A_1 \bowtie_L A_3$ hold and (A_1, A_3) is marginally better or equal than (A_1, A_2), then*

- *$A_1 \succ A_2$ entails $A_1 \succ A_3$,*
- *$A_1 \approx A_2$ entails $A_1 \succ A_3$ or $A_1 \approx A_3$ and*
- *$A_1 \prec A_2$ entails $A_1 \succ A_3$, $A_1 \approx A_3$ or $A_1 \prec A_3$.*

In plain English, if no marginal is worse nor incomparable, then the overall ranking is no worse.

4.2 Further Properties

Properties often discussed in ranking tasks are *transitivity* and *acyclicity*. In general, the body of evidence is such that the marginal comparisons are not transitive – here understood in the sense that if $A_1 >_L A_2$ and $A_2 >_L A_3$ hold, then it can well be the case that one of $A_1 \approx A_3$, $A_1 <_L A_3$, $A_1 \bowtie_L$ or, in exceptional cases, $A_1 \ll_L A_3$ holds.[16] It should hence come as no surprise that the final ranking is, in general, not transitive. Furthermore, it may be the case that the final ranking is cyclic. Intransitive and/or cyclic final rankings can arise even if all marginal comparisons are transitive and acyclic.

Indifference over alternatives is often taken to be an *equivalence relation*. Here, $\perp$ and $\approx$ are by definition symmetric neither is, in general, transitive. Again, this is to be expected in this decision problem with heterogeneous evidence.

Finally, consider a hierarchy and two alternatives A_1, A_2 where there is no evidence for A_1 nor evidence for A_2 on Level $L \geq 2$. We have $A_1 \sim_L A_2$. Now imagine a hypothetical scenario in which the same information is a available but a different hierarchy is used. This new hierarchy is obtained from the old one by ignoring Level L. That is, all information previously categorised as evidence of Level L is now not considered to be evidence at all. In this hypothetical scenario, HiDAD will rank A_1 and A_2 the same way as in the first scenario. This due to the fact, that $\sim_L$ in does not influence aggregated comparisons (Table 5).

In a case where there is a level of evidence in the hierarchy for which there is no evidence available at all one might as well remove this level from the hierarchy.

5 Modifications of HiDAD

There are a number of ways in which HiDAD can be modified to better suit the needs of the DM in a particular decision problem. Firstly, there are different ways in which marginal comparisons could be represented and/or operationalised (modification of Table 1 and Table 2). More nuanced (finer) representations of marginal comparisons are one option, *if* the DM is able and comfortable with making such more nuanced assessments. Invariably, this will lead to a more involved aggregation of marginal comparisons. In case the evidence does not support more nuanced marginal comparisons but only allows for weaker comparisons I would like to recall the problems purely ordinal multi-criterial decision making faces, see Section 2.2.2.

Secondly, (different) aggregated marginal comparisons may have different meanings (modification of Table 3 and Table 4). Thirdly, as already suggested above there is room for different ways of filling the fields marked with * in Table 5 and also other fields in Table 5. For example, the last column of Table 5 devoted to incomparability could be modified to allow lower level evidence to play a larger role.

[16] I think that such cases are *very* rare. However, should the DM assess the evidence thusly, then there have to be good reasons for doing so.

Fourthly, the DM could recommend all non-minimal elements of the final ranking rather than all maximal alternatives or use some other procedure discussed in Section 3.3.

6 Applicability of HiDAD

I now briefly discuss which other problems HiDAD may be applied to and in which cases it does not make sense to apply HiDAD.

6.1 Potential further Areas of Application

In principle, in all other problems which require amalgamation the application of HiDAD may make sense, e.g., judgement aggregation (of experts) where different judgements have different levels of credibility and/or expertise, evidence amalgamation (in general), preference aggregation (in MCDM). There is no reason to think that these other potential areas of application have to be in the medical domain.

6.2 Restricted Scope for Application

There are cases in which the application of HiDAD is not a sensible idea, e.g., when precise quantification is possible or if the importance of the different levels of evidence is judged to vary only incrementally.

I also want to advise against the application of HiDAD in cases where there are less than a handful of levels or more than ten levels. If there are very few evidence levels, then the most important level evidence almost trumps: If $A_i >_1 A_k$ holds, then in only very few cases will A_i not be ranked higher than A_k.

If there are very many levels, then low level evidence is given too much weight in certain instances. For example, for $A_1 >_1 A_2, A_1 \sim_2 A_2, ..., A_1 \sim_{47} A_2, A_1 <_{48} A_2$, $A_1 <_{49} A_2$ and $A_1 <_{50} A_2$ (assuming that the DM has specified 50 levels) HiDAD ranks A_2 higher than A_1. This does not seem right since evidence on Levels 47 – 50 is much less important than evidence on Level 1. The “obvious fix” in this situation is to modify Table 5 by making the entries of the table to also depend on the marginal comparisons of all more important levels and not just their aggregate. I do not endorse any such “fix”.

7 Conclusions

In this chapter, I have put forward a decision aid for ranking problems in medicine which led me to suggest answers to the three questions posed in the introduction: “What constitutes evidence in medicine?”, “How does one amalgamate evidence for medical decision making?” and “How important is the most important kind of evidence?”. Thr answer to the first question was that, for current purposes, evidence is all information which can on its own or jointly with other information conceivably influence the decision problem at hand.[17] The second and third questions were addressed in Section 3.2 and Section 3.3 by the aggregation of marginal comparisons.

I now discuss the claims I base this chapter on and point out some claims I do not make.

7.1 Claims

7.1.1 Normative Claims made

The *normative* claims this approach hinges on are α) all available evidence ought to be taken into account where evidence is to be understood widely (no trumping), β) the most appropriate framework ought to be used to tackle the ranking problem, γ) precise quantification ought to be avoided, if the DM is not able or comfortable with it and δ) overall preferences between alternatives ought to be determined from assessments that are meaningful to the DM.

This approach obeys these claims by α) taking all evidence of all kinds featuring in the specified evidence hierarchy into account and by avoiding trumping, β) the application of MCDM methodology incorporating evidence hierarchies, γ) using a qualitative approach δ) which only relies on qualitative marginal comparisons to construct the final ranking. The marginal comparisons are elicited from the DM by posing simple questions concerning the decision problem with a clear intuitive meaning.

7.1.2 Claims not made

I do not claim that any evidence hierarchy is normatively correct for all decision problems nor do I relativise this claim to the particular decision problem. The choice of an appropriate evidence hierarchy is left to the DM which entails determining the relevant kinds of evidence and ordering these by their importance.

[17] I do think that this definition of evidence could be of use much more generally. I shall here be content with keeping the focus on the discussed ranking problem.

Furthermore, no claims are made that the marginal comparisons or their aggregation are normatively correct. In Section 5, I talked about alternative ways of formalising marginal comparisons and their aggregations.

I do not claim that there exists a quantitative model which outputs the same final ranking as HiDAD.[18] The simplest such quantitative model would use scaling constants reflecting the importance of a criterion, see for example (Billaut et al., 2010, p. 250). To make the use of scaling constants sensible, one has to be clear about the scale (or normalisation procedure) of a criterion otherwise the model makes no sense and is bound to give absurd results. According to (Billaut et al., 2010, p. 250) this non-sensicality is "invariably taught in any basic course on MCDM". I have no idea what a proper scale in the ranking problem is and hence do not entertain the use of a qualitative model using scaling constants. To me, gerrymandering a more complex quantitative model which fits with HiDAD for the only sake of cooking up such a model does not appear like a good use of time.

HiDAD is not a comprehensive decision aid which is fully systematic and free of all human judgement: I did not give any guidelines of how to determine the levels of the evidence hierarchy nor how to rank these levels. Furthermore, there is no suggestion of how exactly to differentiate between, say, $A_k >_L A_i$ and $A_k \sim_L A_i$. Rather, HiDAD is a decision heuristic.

Furthermore, HiDAD does not model all important evidential inferences in medicine. For example, it is here proposed, that items of evidence are put into categories (levels) and that interaction across levels *only* happens via the aggregation of preferences delineated in Table 5. In reality, new pharmacogenomic evidence showing that there are important genetic subgroups within an RCT trial arm, may lead to a reversed interpretation of the trial results and hence a reversed marginal comparison. This kind of inference is aptly depicted in (Clarke et al., 2014, Figure 1).

7.2 Assessing Properties of HiDAD

As with any decision aid, HiDAD is not universally applicable, see Section 6. While a restricted applicability is surely not ideal, I think that the kind of decision problems HiDAD can be applied to is important enough to warrant this approach.

As already indicated by its name, I think that compatibility (Section 4.1) is desirable. Failure of acyclicity and full equivalence of the indifference relations $\perp$ and $\approx$ (Section 4.2) cannot be blamed on HiDAD. Rather, it is a consequence of the DM's less-than-godlike evidential situation.

HiDAD is hierarchy neutral, in that HiDAD is not tied to any particular hierarchical ordering of the levels of evidence and may be applied whatever the DM deems

[18] If precise quantification were possible, then I do recommend to use these numbers. However, I supposed that precise quantification is not possible and hence went down a qualitative path. For cases in which precise quantification of the importance of criteria is feasible the reader is referred to (Mussen et al., 2009; Tervonen et al., 2011).

the appropriate hierarchy at the time. I take this neutrality to be desirable. Those worried about the under-appreciation of "lower level" evidence have two ways of incorporating their thinking in HiDAD: 1) Re-arrange the (ordering of the) levels of the evidence hierarchy used. 2) On the more important levels, raise the bar for setting $A_k \gg_L A_i$.

HiDAD demands the incorporation of all evidence in a systematic way, as long as all items of evidence are of one kind which is ranked in the hierarchy of levels of evidence. Supposing that the DM can (be aided to) specify such a hierarchy, this requirement of total evidence ought to be a corner stone in evidence based medicine. I hence think that this requirement is desirable.

HiDAD outputs the final ranking which is completely determined by a fixed hierarchy of levels of evidence, the marginal comparisons and their aggregations which all are, relatively, easily articulated. This makes the entire decision making process transparent, explicit and reproducible by erasing all further room for subjective choice and arbitrariness.

Unlike GRADE, the operationalisation of marginal comparisons in HiDAD is in terms of a simple questions with intuitive implications for the decision problem. Furthermore, the number of judgements required to determine the final ranking of two alternatives given the body of available evidence and the evidence levels and their importance equals the number of levels of evidence. Since this number is (roughly) between five and ten I think that the final ranking of a pair of alternatives is not swamped by the number of human judgements.

To me, it is desirable that levels for which there is no evidence can be ignored.

Finally, I do not take issue with decision heuristics. Gerd Gigerenzer, for example, has long argued that decision heuristics make a good deal of sense and often outperform the most sophisticated approaches, the reader finds his arguments in, e.g., (Gigerenzer and Gaissmaier, 2011; Marewski and Gigerenzer, 2012). Decision heuristics, such as GRADE and HiDAD aim at supporting decision making processes in the actual world; they were designed with the intention to outperform intuitive decisions or other formal methods. In the design process there are a number design choices to be made (GRADE: number of up- and down-grades possible, when to up- or down-grade, HiDAD: choice of marginal comparisons, their aggregations). There is no normatively correct way of making these design choices. The outputs of GRADE and HiDAD are hence not normatively correct in a strong sense.

Whether or not HiDAD is successful cannot be assessed based on purely academic consideration, it will depend on the experiences of the stakeholders and – most importantly – the patients who are treated by taking the guidelines produced by applying HiDAD into account.

Acknowledgements The idea for this chapter arose when the author was a research assistant on a project on the "Optimal design of biofuel production by microalgae" at INRA, UR0050, Laboratoire de Biotechnologie de l'Environnement (2009 - 2010). Progress all but ceased when the author joined the "From objective Bayesian epistemology to inductive logic" AHRC-funded project at the University of Kent. The great majority of the work was carried out after the author joined the ERC-funded project "Philosophy of Pharmacology: Safety, Statistical Standards, and Evidence Amalgamation" (grant 639276) at the LMU Munich. Regarding this chapter, the author benefited from a number of discussions with Seamus Bradley, Ricardo Büttner, Teddy Groves, Laurent Lardon, Barbara Osimani, Roland Poellinger, David Teira and Jon Williamson as well as the members of the Environmental Lifecycle and Sustainability Assessment group. He would also like to thank an anonymous referee for the European Journal of Operational Research and three anonymous reviewers for this volume as well as the editors of this volume for their thoughtful comments and insights.

References

Alonso-Coello, P., Schünemann, H. J., Moberg, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., Treweek, S., Mustafa, R. A., Rada, G., Rosenbaum, S., Morelli, A., Guyatt, G. H., and Oxman, A. D. (2016). GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ*, 353.

Aumann, R. J. (1962). Utility Theory without the Completeness Axiom. *Econometrica*, 30(3):445–462.

Belton, V. and Stewart, T. J. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer.

Bertamini, M. and Munafó, M. R. (2012). Bite-Size Science and Its Undesired Side Effects. *Perspectives on Psychological Science*, 7(1):67–71.

Bes-Rastrollo, M., Schulze, M. B., Ruiz-Canela, M., and Martinez-Gonzalez, M. A. (2013). Financial Conflicts of Interest and Reporting Bias Regarding the Association between Sugar-Sweetened Beverages and Weight Gain: A Systematic Review of Systematic Reviews. *PLOS Medicine*, 10(12):1–9.

Billaut, J.-C., Bouyssou, D., and Vincke, P. (2010). Should you believe in the Shanghai ranking? An MCDM view. *Scientometrics*, 84:237–263.

Borm, G. F., Lemmers, O., Fransen, J., and Donders, R. (2009). The evidence provided by a single trial is less reliable than its statistical analysis suggests. *Journal of Clinical Epidemiology*, 62(7):711–715.

Boström, H., Andler, S. F., Brohede, M., and Johansson, R. (2007). On the Definition of Information Fusion as a Field of Research. Technical report, University of Skövde.

Boutilier, C. (1994). Toward a logic for qualitative decision theory. In *Proceedings of KR*, volume 94, pages 75–86. Morgan Kaufmann.

Bouyssou, D., Marchant, T., Pirlot, M., Tsoukiàs, A., and Vincke, P. (2006). *Evaluation and Decision Models with Multiple Criteria*. Springer.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14:365–376.

Canadian Task Force on the Periodic Health Examination (1979). The periodic health examination. *Canadian Medical Association Journal*, 121(9):1193–1254.

Carnap, R. (1947). On the Application of Inductive Logic. *Philosophy and Phenomenological Research*, 8(1):133–148.

Cartwright, N. (2007). Are RCTs the Gold Standard? *Biosocieties*, 2(1):11–20.

Cartwright, N. and Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice*, 16(2):260–266.

Chalmers, I., Hedges, L. V., and Cooper, H. (2002). A Brief History of Research Synthesis. *Evaluation & the Health Professions*, 25(1):12–37.

Chan, A.-W. and Altman, D. G. (2005). Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet*, 365(9465):1159–1162.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine*, 57(6):745–747.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2014). Mechanisms and the Evidence Hierarchy. *Topoi*, 33:339–360.

Cooper, N., Coyle, D., Abrams, K., Mugford, M., and Sutton, A. (2005). Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *Journal of Health Services Research & Policy*, 10(4):245–250.

Doll, R. and Peto, R. (1980). Randomised controlled trials and retrospective controls. *British Medical Journal*, 280:44.

Doyle, J. and Thomason, R. H. (1999). Background to qualitative decision theory. *AI Magazine*, 20(2):55–68.

Dubois, D., Fargier, H., and Perny, P. (2002). On the Limitations of Ordinal Approaches to Decision-making. In Fensel, D., Giunchiglia, F., McGuinness, D. L., and Williams, M.-A., editors, *Proceedings of KR*, pages 133–146. Morgan Kaufmann.

Dubois, D., Fargier, H., and Prade, H. (1997). Decision-making Under Ordinal Preferences and Comparative Uncertainty. In *Proceeding of UAI*, pages 157–164.

Dubois, D., Fargier, H., Prade, H., and Sabadin, R. (2009). A survey of qualitative decision rules under uncertainty. In *Decision-making Process*, chapter 11, pages 435–473. Wiley Online Library.

Earman, J. (1992). *Bayes or Bust?* MIT Press.

Etz, A. and Vandekerckhove, J. (2016). A Bayesian Perspective on the Reproducibility Project: Psychology. *PLoS ONE*, 11(2).

Evans, D. (2003). Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, 12(1):77–84.

Every-Palmer, S. and Howick, J. (2014). How evidence-based medicine is failing due to biased trials and selective publication. *Journal of Evaluation in Clinical Practice*, 20(6):908–914.

Figueira, J., Greco, S., and Ehrgott, M. (2005). *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, 62(1):451–482.

GRADE Working Group (2004). Grading quality of evidence and strength of recommendations. *British Medical Journal*, 328(7454):1490–1494.

Guyatt, G., Oxman, A. D., Sultan, S., Brozek, J., Glasziou, P., Alonso-Coello, P., Atkins, D., Kunz, R., Montori, V., Jaeschke, R., Rind, D., Dahm, P., Akl, E. A., Meerpohl, J., Vist, G., Berliner, E., Norris, S., Falck-Ytter, Y., and Schünemann,

H. J. (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(2):151–157.

Guyatt, G. H., Oxman, A. D., Schünemann, H. J., Tugwell, P., and Knottnerus, A. (2011). GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology*, 64:380–382.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, 336(7650):924–926.

Horton, R. (2004). Vioxx, the implosion of Merck, and aftershocks at the FDA. *The Lancet*, 364(9450):1995–1996.

Howick, J. and co workers (2011). The Oxford 2011 Levels of Evidence.

Howick, J. H. (2011). *The Philosophy of Evidence-Based Medicine*. Blackwell.

Jüni, P., Nartey, L., Reichenbach, S., Sterchi, R., Dieppe, P. A., and Egger, M. (2004). Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *The Lancet*, 364(9450):2021–2029.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with Multiple Objectives*. Cambridge Books. Cambridge University Press.

Kelly, M. P. and Moore, T. A. (2012). The judgement process in evidence-based medicine and health technology assessment. *Social Theory & Health*, 10(1):1–19.

Kelly, T. (2015). Evidence. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition.

Khangura, S., Polisena, J., Clifford, T. J., Farrah, K., and Kamel, C. (2014). Rapid review: an emerging approach to evidence synthesis in health technology assessment. *International Journal of Technology Assessment in Health Care*, 30(1):1–8.

Krumholz, H. M., Ross, J. S., Presler, A. H., and Egilman, D. S. (2007). What have we learnt from Vioxx? *British Medical Journal*, 334(7585):120–123.

La Caze, A. (2009). Evidence-Based Medicine Must Be ... *Journal of Medicine and Philosophy*, 34(5):509–527.

Landes, J., Osimani, B., and Poellinger, R. (2017). Epistemology of Causal Inference in Pharmacology. *European Journal for Philosophy of Science*. 47 pages.

Lehtinen, A. (2013). On the Impossibility of Amalgamating Evidence. *Journal for General Philosophy of Science*, 44(1):101–110.

Marewski, J. N. and Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, 14:77–89.

McGauran, N., Wieseler, B., Kreis, J., Schuler, Y.-B., Kolsch, H., and Kaiser, T. (2010). Reporting bias in medical research - a narrative review. *Trials*, 11(37):1–15.

Mussen, F., Salek, S., and Walker, S. (2009). *Benefit-Risk Appraisal of Medicines*. John Wiley & Sons.

Onakpoya, I. J., Heneghan, C. J., and Aronson, J. K. (2016). Worldwide withdrawal of medicinal products because of adverse drug reactions: a systematic review and analysis. *Critical Reviews in Toxicology*, 0(0):1–13.

Osimani, B. (2014a). Hunting Side Effects and Explaining Them: Should We Reverse Evidence Hierarchies Upside Down? *Topoi*, 33(2):295–312.

Osimani, B. (2014b). Safety vs. efficacy assessment of pharmaceuticals: Epistemological rationales and methods. *Preventive Medicine Reports*, 1:9–13.

Price, K. L., Amy Xia, H., Lakshminarayanan, M., Madigan, D., Manner, D., Scott, J., Stamey, J. D., and Thompson, L. (2014). Bayesian methods for design and analysis of safety trials. *Pharmaceutical Statistics*, 13(1):13–24.

Reiss, J. (2015). A pragmatist theory of evidence. *Philosophy of Science*, 82(3):341–362.

Revicki, D. A. and Frank, L. (1999). Pharmacoeconomic Evaluation in the Real World. *PharmacoEconomics*, 15(5):423–434.

Russo, F. and Williamson, J. (2007). Interpreting Causality in the Health Sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Sackett, D. L., Straus, S. E., Richardson, W. S., and Haynes, R. B. (2000). *Evidence-based medicine: how to practice and teach EBM*. Churchill Livingstone, 2 edition.

Savage, L. J. (1954). *The Foundations of Statistics*. Dover Publications.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Shafer, G. and Srivastava, R. (1990). The Bayesian and Belief-Function Formalisms: A General Perspective for Auditing. *Auditing: A Journal of Practice & Theory*, 9:110–137.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Solomon, M. (2011). Just a paradigm: evidence-based medicine in epistemological context. *European Journal for Philosophy of Science*, 1(3):451–466.

Stegenga, J. (2013). An impossibility theorem for amalgamating evidence. *Synthese*, 190(12):2391–2411.

Stegenga, J. (2014). Down with the Hierarchies. *Topoi*, 33(2):313–322.

Tan, S.-W. and Pearl, J. (1994). Qualitative Decision Theory. In *Proceedings of AAAI*, volume 2, pages 928–933.

Tervonen, T., van Valkenhoef, G., Buskens, E., Hillege, H. L., and Postmus, D. (2011). A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine*, 30(12):1419–1428.

Thomas, J., Newman, M., and Oliver, S. (2013). Rapid evidence assessments of research to inform social policy: taking stock and moving forward. *Evidence & Policy: A Journal of Research, Debate and Practice*, 9(1):5–27.

Troffaes, M. C. M. and de Cooman, G. (2014). *Lower Previsions*. Wiley.

Twinning, W. (2011). Moving Beyond Law: Interdisciplinarity and the Study of Evidence. In Phil Dawid, William Twinning, M. V., editor, *Evidence, Inference and Enquiry*, chapter 4, pages 73–118. OUP.

Upshur, R. (1995). Looking for Rules in a World of Exceptions: reflections on evidence-based practice. *Perspectives in Biology and Medicine*, 48(4):477–489.

Urfalino, P. (2012). Reasons and Preferences in Medicine Evaluation Committees. In Landemore, H. and Elster, J., editors, *Collective Wisdom*, pages 173–202. Cambridge University Press.

van Valkenhoef, G., Tervonen, T., Zwinkels, T., de Brock, B., and Hillege, H. (2013). ADDIS: A decision support system for evidence-based medicine. *Decision Support Systems*, 55(2):459–475.

Vandenbroucke, J. P. (2008). Observational Research, Randomised Trials, and Two Views of Medical Science. *PLoS Medicine*, 5(3):e67.

Williamson, J. (2015). Deliberation, Judgement and the Nature of Evidence. *Economics and Philosophy*, 31:27–65.

Worrall, J. (2002). What Evidence in Evidence-Based Medicine? *Philosophy of Science*, 69(3):316–330.

Worrall, J. (2007a). Evidence in Medicine and Evidence-Based Medicine. *Philosophy Compass*, 2(6):981–1022.

Worrall, J. (2007b). Why There's no Cause to Randomize. *British Journal for the Philosophy of Science*, 58(3):451–88.

Worrall, J. (2010). Evidence: philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice*, 16(2):356–362.