Synthese (2019) 196:3083–3102
https://doi.org/10.1007/s11229-017-1627-1

S.I.: EVIDENCE AMALGAMATION IN THE SCIENCES

The problem of evaluating automated large-scale evidence aggregators

Nicolas Wüthrich[1] · **Katie Steele**[2]

Received: 7 November 2016 / Accepted: 13 November 2017 / Published online: 28 November 2017

Abstract In the biomedical context, policy makers face a large amount of potentially discordant evidence from different sources. This prompts the question of how this evidence should be aggregated in the interests of best-informed policy recommendations. The starting point of our discussion is Hunter and Williams' recent work on an automated aggregation method for medical evidence. Our negative claim is that it is far from clear what the relevant criteria for evaluating an evidence aggregator of this sort are. What is the appropriate balance between explicitly coded algorithms and implicit reasoning involved, for instance, in the packaging of input evidence? In short: What is the optimal degree of 'automation'? On the positive side: We propose the ability to perform an adequate robustness analysis (which depends on the nature of the input variables and parameters of the aggregator) as the focal criterion, primarily because it directs efforts to what is most important, namely, the structure of the algorithm and the appropriate extent of automation. Moreover, where there are resource constraints on the aggregation process, one must also consider what balance between volume of evidence and accuracy in the treatment of individual evidence best facilitates inference. There is no prerogative to aggregate the total evidence available if this would in fact reduce overall accuracy.

✉ Nicolas Wüthrich
N.Wuethrich@lse.ac.uk

Katie Steele
katie.steele@anu.edu.au

[1] Department for Philosophy, Logic, and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK

[2] School of Philosophy, Australian National University, HC Coombs Building, Canberra 0200, Australia

Keywords Evidence aggregation · Evidence-based medicine · Statistical meta-analysis · Robustness analysis

1 Introduction

It is now a commonplace that policy-making should be evidence-based (Cartwright and Hardie 2012; Montuschi 2009): for example, class size regulations, housing policies, and dosage recommendations regarding new substances, all should be designed, as far as practicable, in light of the best available evidence. Policy makers demand evidence for many reasons, including assessing the potential impact of planned interventions and providing greater accountability to relevant stakeholders. However, in these social and biomedical contexts, policy makers are facing an enormous amount of evidence (Stegenga 2011, p. 497; Krimsky 2005, p. 129; Weed 2005, p. 1545). This prompts the question of how these large sets of diverse and potentially conflicting evidence should, if at all, be aggregated to facilitate best-informed policy recommendations.

In this paper, we focus on this problem in the medical context. Evidence aggregation in medicine can involve a host of data from different sources, such as observational studies, randomized controlled trials, meta-analyses, and expert judgment, which often confirm conflicting hypotheses. Throughout this paper we refer to aggregation procedures involving evidential input of this sort, i.e., high volume, highly diverse and potentially conflicting evidence, as 'large-scale' evidence aggregation.

In a series of recent papers, Hunter and Williams (2010, 2012, 2013) have developed a comprehensive decision support tool for medical policy makers. Their framework motivates our discussion here, because it is ambitious and striking and thus brings to light a number of general questions. First, their approach is large-scale and highly computational—it involves an explicit algorithm, such that, once initiated, relatively little user input is required. (Throughout the paper, we refer to their approach as 'highly automated'.) Second, they provide a full account of the steps to choosing a (medical or treatment) policy. That is, their framework or algorithm involves both an epistemic component (assessing hypotheses concerning the various causal effects of the relevant treatments) and a preference-driven component (weighing the treatments in light of these hypotheses). This paper focuses on the epistemic component, which is in line with what many consider to be the problem of 'evidence aggregation', but we will also briefly address how this component fits into the bigger treatment-selection picture. As regards the epistemology alone, Hunter and Williams' proposal invites reflection about relatively under-appreciated issues: Does it make sense to deliver a highly automated large-scale evidence aggregator? How do we assess the optimal extent of automation for any given type of evidence aggregation task?

Let us be clear what we take automation to consist in and what the alternative to automation is. We propose that evidence aggregators are automated to the extent that they explicitly encode procedures or algorithms that are fixed, transparent, and dependent on inputs characterised in a particular way. (Ideally these procedures would also be encoded in such a way that they can be run on a computer at high speed, and also so that values for key parameters can be easily changed. Here we will assume that 'explicit algorithms' are implemented wisely so that they allow for these benefits. In

this sense, automation implies computability in the discussion that follows.) The non-automated aspects of aggregation involve implicit reasoning that affords flexibility, in the sense that how inputs are interpreted and weighed may be determined 'on the fly'. It is effectively expert judgment that may or may not be capable of explicit articulation. So, the main contrast is between reasoning that is made explicit or public, and reasoning that is 'behind the scenes', a 'black box' aspect of the procedure, so to speak. The more the packaging of the input evidence relies on 'behind the scenes' reasoning, for instance, the less automated the aggregator.

We defend two claims. First, a negative claim: We argue that the above questions are non-trivial. It is very difficult to specify criteria for determining the appropriate extent of automation for a large-scale evidence aggregator, such as Hunter and Williams', even when it comes to assessing a single, well-specified causal hypothesis. This problem deserves more philosophical attention. Second, on a more positive note: Despite this lack of clarity, we defend *one* criterion for assessing the reliability of large-scale evidence aggregators: the capacity, given the nature of the input variables and parameters of the aggregator, to perform an *adequate* robustness analysis. In effect, this criterion directs attention to the question that matters: Does the aggregation function have appropriate parameters and associated input variables (even if there is uncertainty about the precise parameter values)? We go on to argue that resource constraints complicate the assessment of what algorithm structure enhances reliability or sound inference. In this case volume of evidence competes with accuracy in processing individual evidence. There is no prerogative to process the total evidence available via a relatively crude algorithm, if this would in fact reduce overall accuracy.

The paper is structured as follows. First, we sketch the general problem of evidence-based comparison of policy (or treatment) options, our focus being the epistemic or evidential component. The goal here is to see how far standard logic(s) of inference constrain or guide the task of constructing a large-scale evidence aggregator (Sect. 2). While formal logic does assist in framing the aggregation problem, it does not settle the substantial questions of interest to us concerning the optimal extent of automation. This should be even more apparent from the description of Hunter and Williams' approach (Sect. 3). So we need to develop further criteria for assessing a large-scale evidence aggregator. We go on to introduce the ability to perform an adequate robustness analysis as at least one such evaluation criterion, and we compare the case of unlimited resources (e.g. computation time) to the case of a constrained resource budget (Sect. 4). Finally, we defend our criterion against some objections (Sect. 5).

2 Evidence-based decisions: separating the key components

Hunter and Williams aim to provide answers to what is ultimately of interest in practical medical contexts: their decision support tool outputs claims like 'treatment 1 is medically better than treatment 2 for patients of type X'. While such claims appear to concern only medical 'facts', they are ultimately claims about preferences—whether treatment 1 is *preferred*, on all-things-considered medical grounds, to treatment 2.[1]

[1] Of course, non-medical considerations (prominently, monetary cost) may also play a role in advice relating to the comparison of medical treatments, but we leave such considerations aside.

Table 1 A simplified decision problem between treatments 1 and 2 in light of the two criteria 'pain relief' and 'blood pressure'

	Pain relief	Blood pressure
Treatment 1	Pain 1	Blood 1
Treatment 2	Pain 2	Blood 2

Thus these sorts of claims are *deceptively* simple; the first thing to appreciate is that they involve both a factual and an evaluative element.

An example helps to illustrate what is at stake: Consider a doctor deciding what treatment to recommend a patient. Assume there are just two treatments. (Note that, for the most part, we will assume, for simplicity, that there are just two options to choose between.) We will further simplify the problem by assuming that the doctor has a good idea already of the characteristics of the patient. That is, we are not considering the problem of diagnosis, which is a tricky problem in its own right. Assume that in the case at hand both the impact on pain relief and the impact on blood pressure matter. Representing the situation in terms of a 'multi-criteria' decision problem, as per Table 1, helps to illuminate the dimensions of the problem.

Recall that we ultimately want to assess the claim: 'treatment 1 is more choice-worthy (i.e., medically better, with respect to pain relief and blood pressure), relative to treatment 2, for patients of type X'. Table 1 makes explicit that this involves comparing the treatments with respect to both pain relief and blood pressure, and then ultimately ranking the treatments, depending on their performance in these respects, and the relative importance of the two criteria. So the decision problem at hand involves both facts (the causal effects of the treatment options with respect to pain and blood pressure) and values (the relative importance of differences in pain relief and blood pressure).

Appreciation of the basic decision model is the first step in analysing Hunter and Williams' proposal, which we outline in Sect. 3. It allows us to distinguish the various components of their decision support tool. The aim in the remainder of this section is to consider how much further standard logic can take us in analysing Hunter and Williams' (and other comparable) proposal(s). Our particular interest is the *evidence aggregation problem* (in s2.1–s2.2 below); we mention the literature on multi-criteria decision-making only briefly at the end of the section (in s2.3). One might hope that standard logic of inference is the key to assessing a highly automated evidence aggregator such as devised by Hunter and Williams. We suggest, however, that one can construe the logic of evidence aggregation in a variety of ways; in any case, the bare logic does not in itself settle the more substantial questions concerning optimal design. These questions are taken up in later sections.

2.1 The evidence aggregation problem

Recall that, for simplicity, we restrict our attention to the comparison of just two treatments: In such a case, the key empirical questions are whether (and, ideally, by how much) one treatment is better than the other on each respective medical dimension (e.g. pain relief). That is, we want to know the sign and ideally also the magnitude

of the difference in the relevant *effect size(s)* for the two treatments.[2] This question can be translated into a set of hypotheses: either very fine-grained hypotheses concerning the precise difference in effect size between the two treatments, or else more coarse-grained hypotheses concerning simply whether or not treatment 1 is better than treatment 2 along the dimension in question.

The problem of evidence aggregation is about what the total evidence at hand says about any given target partition of hypotheses. For the Bayesian, there is a straightforward, albeit abstract, answer to this question: The *impact* of the evidence on the relative credibility (measured in terms of probabilities) of two hypotheses depends on the relative *likelihoods* for the hypotheses, or in other words, the relative probability of the total evidence conditional on the respective hypotheses. That is, evidential impact depends on the *likelihood ratio*, stated here in terms of the probability function Pr, for two hypotheses H and H' given total evidence amounting to the conjunction of evidence propositions $E_1, \ldots, E_n$:

$$\frac{\Pr(E_1, E_2, \ldots, E_n|H)}{\Pr(E_1, E_2, \ldots, E_n|H')}$$

In the case that $E_1, E_2, \ldots, E_n$ are conditionally independent with respect to H, then the likelihood for the total evidence is just the product of their separate likelihoods: $\Pr(E_1, E_2, \ldots, E_n|H) = \Pr(E_1|H) \times \Pr(E_2|H) \times \ldots \times \Pr(E_n|H)$. In the case of dependencies between the evidence propositions, however, the likelihoods cannot be thus decomposed.

Bayesian logic does not itself dictate the content of the evidence propositions, denoted here $E_1, \ldots, E_n$. But constraints assist with tractability. Many applications (including those of interest here) are suited to the following 'witness model' constraint: The individual 'pieces of evidence', $E_1, \ldots, E_n$, take the form of *witness reports*, that is, direct reports on the truth of the hypotheses of interest (whether a probability of truth or a binary assessment of truth). The witness reports are furthermore taken to be independent.[3] The likelihood ratio associated with any such report is effectively a measure of the reliability of the witness; it is the ratio of the probability that the witness reports that the hypothesis is true given that it is in fact true to the probability that the witness reports that the hypothesis is true given that it is false. The challenge of applying the witness schema in any given case is to identify what are the separate pieces of evidence that can count as independent witness reports. These reports must be subsequently assessed for reliability.

The Bayesian witness model has much in common with models for aggregating individual opinions to arrive at a collective opinion (see List 2012 for an overview of the aggregation of a set of logically connected binary opinions and Genest and Zidek 1986 for an overview of the aggregation of probabilistic opinions). Both kinds of

[2] The pain relief effect size, for instance, might be characterised in terms of the proportion of subjects who report significant decrease in pain after treatment.

[3] The independence requirement of these models is generally only approximately respected in practice. Salient witness reports are simply treated as independent, whether or not there are doubts about this. We tend to go along with this stance in what follows. While assuming independence without warrant may compromise one's inferences/judgments, this is not our focus in this paper.

model involve assessments of the *reliability* of the individual 'witness' opinions, and they assume that (or are more tractable to the extent that) the individual opinions are independent of each other. The collective aggregation approach is more straightforward in that the final assessment of the hypotheses is a direct aggregate of the actual witness reports.[4] Indeed the latter model is perhaps the best way to interpret Hunter and Williams' proposal in the next section. Either way, we will see below (and also in the next section) that the bare logic of a 'witness model' does not greatly constrain the evidence aggregation task.

2.2 The literature on evidence aggregation vis-à-vis the schema

The existing literature on aggregating evidence for causal hypotheses accords well with the witness schema. One might interpret the key questions as follows: How should individual 'witness reports' be delineated to (at least approximately) preserve *independence* and how should the reliability of these reports be assessed? What algorithm should be employed for aggregating the reported causal conclusions? Consider an example report amounting to a statistical reject/accept result concerning some causal claim. This is effectively a witness report that has some degree of reliability. Indeed, there is a preoccupation in the literature (in particular, the 'hierarchy of evidence' literature associated with the 'evidence-based medicine/policy' movement; see Brendan et al. 2014 for an overview) with how to assess the reliability of study results given the design in question (the sample size, the controls for bias, etc.), and furthermore with how to aggregate the results, for instance, whether supposedly more reliable study types should 'trump' others or whether they should all play a role in determining the final causal conclusions.[5]

Most evidence hierarchies place *statistical meta-analyses* near the top.[6] One might worry that the witness model does not account for this kind of conglomerate evidence, where the effect size measures from different studies (perhaps the difference in efficacy of two treatments) are combined in sophisticated ways. By way of response: There are two ways one might reconcile statistical meta-analysis with the witness schema. The first is to see meta-analysis as an instance of the general schema, with a particular choice of algorithm for aggregating the witness reports, and limited to a certain kind

[4] In the Bayesian setting, the witness reports are mere indicators of the truth.

[5] For instance, the hierarchy introduced by the UK National Institute for Health and Care Excellence (NICE 2006) places meta-analyses, systematic reviews of randomised controlled trials (RCTs), and RCTs above case-control or cohort studies, non-analytical studies such as case reports, and expert opinion. In addition to worries about undue privileging of RCTS, a major criticism of the evidence hierarchies is that mechanistic evidence and expert opinion are distinguished from correlational evidence and relegated to the bottom of the hierarchy (see Brendan et al. 2014). The claim is that mechanistic evidence for causation is rather complementary to statistical evidence (see the Russo-Williamson thesis, Russo and Williamson 2007). This criticism does not necessarily undermine the witness schema. Instead, one can draw a lesson of caution from this debate regarding what counts as independent witness-style evidence. In particular, one might contend that evidence of mechanisms and evidence of probabilistic-association cannot be treated as separate 'witnesses'.

[6] As per the NICE hierarchy (see previous footnote).

of evidence (generally the fine-grained effect-size results of RCTs).[7] The second is to treat an individual meta-analysis as itself a single study or witness report, perhaps a very reliable one if it is based on considerable data from multiple experiments (but see Stegenga 2011 for doubts about the reliability of meta-analyses). The latter route is attractive because in practice there are many meta-analyses at hand that account for some, but not all, of the raw statistical data that could serve as potential input for assessing a causal hypothesis.[8]

In general, the above discussion underscores a point made above, namely that the witness schema, while offering some useful constraints, is nonetheless a very flexible model for evidence aggregation. Clearly a lot depends on how the evidence is divided into independent witnesses and assessed for reliability. That is, it is apparent that a lot of the difficult questions in negotiating evidence are shifted to the 'pre-processing' of the evidence, and away from the final aggregation task. So, the general dictates of logic—here spelled out in terms of a witness model—can provide a basic framework for evidence aggregation, but further substantial aspects of the process are inevitably left unspecified. That is not to say that these further details of identifying, weighing and aggregating evidence cannot in large part be automated via an explicit algorithm. How much automation is a good idea is of course the question that remains unanswered.

2.3 Return to the wider multi-criteria decision problem

As noted earlier, in order to finally make a choice of treatment, one not only has to tackle the above evidence aggregation problems for the relevant empirical hypotheses, but also the *problem of combining, according to decision theoretic principles, the various empirical and evaluative claims that are pertinent to the final choice of option*. This is a hugely contested issue and falls under the literature of 'multi-criteria decision analysis' (for an overview see Keeney and Raiffa 1993). There are a host of suggested methods for making a choice in the face of options that are ranked differently according to different criteria which themselves have differing importance for the decision-maker. An important specification of these methods is whether the ranking of options for each dimension can be represented in cardinal terms or rather only in ordinal terms: i.e., can one specify how much better/worse treatment 1 is on pain relief compared to treatment 2 for patient type X or can one only say which is better/worse? When cardinal information is plausibly available and the criteria are comparable in cardinal terms, the decision analysis can follow standard Bayesian principles in the form of

[7] Stegenga's (2011, pp. 497–498) description of the steps involved in meta-analysis fits with a version of the witness aggregation schema: (a) selecting the primary studies, (b) determining the appropriate outcome measure for each study (such as effect size), (c) weighing each study (usually according to its size and quality, for example by using the inverse of the variance of the effect estimate), and (d) calculating the weighted average of the effect sizes across the studies.

[8] Let us mention also more complex studies that involve inferring a causal graph from a joint probability distribution over a number of variables, such as Bayesian nets methods (for an introduction, see Ben-Gal 2007). Here again we might treat any such study as a single witness report with respect to the particular cause and effect that is of interest. While our evidence aggregation problem is 'large scale' when it comes to the diversity of evidence, it is in another sense simple in that we restrict attention to a single cause and effect relationship, as opposed to a more detailed and complex causal network.

expected utility theory. When only ordinal information is available, the controversies regarding how to make an overall choice are more extensive.

3 A striking proposal: Hunter and Williams' highly automated aggregator

Hunter and Williams claim to offer "a new framework for representing and synthesizing knowledge from clinical trials involving multiple outcome indicators." (Hunter and Williams 2012, p. 1) We summarize what we take to be the key tenets of their account. The upshot is that their procedure illustrates the main claim made in the previous section: to determine the optimal extent of automation, further criteria are needed over and above the information provided by referencing standard logic(s) of inference.[9] The goal of Hunter and Williams' approach is to come up with an ordinal ranking of two treatment options for a specific patient (2013, p. 16).[10] For example: Is the treatment of contraceptive pill or no treatment better, when we look at pregnancy, ovarian cancer, and breast cancer as outcome indicators for patient Y? (ibid., p. 13).[11]

Their approach contains two aggregation elements:

1. The aggregation of evidence for each outcome indicator, which itself involves (a) delineating evidence, (b) assessing reliability of evidence, and (c) an aggregation rule.
2. The overall multi-criteria problem, taking into account all relevant outcome indicators.

Given the focus of our discussion, we explain the first of these two aggregation steps in more detail. We briefly sketch their approach to the second step for completeness.

To aggregate the available evidence regarding the various outcome indicators, Hunter and Williams introduce an evidence table. This table delineates evidence in terms of different studies. For each study, a number of key input variables are filled in or evaluated (ibid., pp. 6–7):

(a) the pair of treatments (e.g. contraceptive pill, no treatment),
(b) the outcome indicator denoting the dimension along which the treatments are compared (e.g. breast cancer),
(c) the value of the outcome indicator given a particular measure. If one adopts the relative risk measure, then the ratio between the portion of people displaying the outcome given treatment 1 (e.g. breast cancer given contraceptive pill) and the portion of people displaying the outcome given treatment 2 (e.g. breast cancer given no treatment) is calculated,

[9] Interestingly, Hunter and Williams themselves point out the need of some sort of normative standard to assess their framework when they compare their results with the recommendations in the NICE guidelines (Hunter and Williams 2012, Section 10).

[10] We discuss solely the case involving two treatment options. Hunter and Williams' proposal is able to deal with multiple pairwise comparisons of treatments.

[11] Here Y refers to the patient class in which the individual patient falls (see Hunter and Williams 2010, p. 119).

(d) the net outcome indicating whether T_1 is superior, inferior, or equal to T_2. The net outcome is determined by the measured value of the outcome indicator and whether the outcome indicator is desirable for a patient class. For example, given a value of 1.04 (assuming the above specified measure of relative risk and the outcome indicator 'breast cancer'), contraceptive pill is inferior to no treatment (ibid., p. 13),
(e) whether the measured value of the outcome indicator is statistically significant, and
(f) the evidence type (e.g. RCT study, cohort study, meta-analysis, or network analysis).[12]

These input variables effectively inform what the evidence says about the treatments and how reliable it is. The information in the evidence table is effectively the total available input for the automated aggregation procedure.

Now, the different studies in the evidence table are treated as independent witnesses. For each witness, its reliability is determined via so-called *meta-arguments*. Hunter and Williams consider the following meta-arguments (for identifying unreliable evidence): 'the evidence contains flawed RCTs', 'the evidence contains results that are not statistically significant', 'the evidence is from trials that are for a very narrow patient class', 'the evidence has outcomes that are not consistent'. These meta-arguments return binary 'in or out' results (meaning that a study either feeds into the aggregation function or it does not) for computational ease and theoretical simplicity (Hunter and Williams 2012, pp. 4, 9, 20).

With these elements in place, we can discuss Hunter and Williams' aggregation rule for any given outcome indicator. It is based on the notion of an *inductive argument*, where this is a pair $(X, \epsilon)_i$, with X being a subset of evidence and ϵ the claim that either T_1 is superior to T_2, T_1 is inferior to T_2, or T_1 is equal to T_2 (ibid. pp. 3–4, 11) with respect to outcome indicator i.[13] The set of inductive arguments is constructed in the following way. To start, the available evidence (concerning T_1 and T_2) is divided into three subsets SUPERIOR, INFERIOR, and EQUIVALENT. The subset SUPERIOR contains all rows of the evidence table (i.e. evidence pieces) for which T_1 was shown to be superior to T_2. The subsets INFERIOR and EQUITABLE are defined via the inferiority and equivalence relations, respectively (ibid., pp. 10–11). The inductive arguments are determined by permissible inference rules from the set of evidence. Hunter and Williams propose the following three simple inference rules that apply to cases where evidence is not conflicting (assuming that X is a subset of evidence, ibid. 11):

1. If $X \subseteq$ SUPERIOR, then $T_1 > T_2$
2. If $X \subseteq$ INFERIOR, then $T_1 < T_2$
3. If $X \subseteq$ EQUITABLE, then $T_1 \sim T_2$

[12] This is merely one candidate list for characterising evidence. Hunter and Williams note that when applying their framework further information should be captured by the evidence table if practicable; for example, the sample size of a trial, the geographical location for each trial, the drop-out rate, the method of randomization, and whether a trial used a narrow patient class (Hunter and Williams 2012, p. 7).

[13] We have augmented Hunter and Williams' notation for an inductive argument to indicate that they are indexed to outcome indicators.

For example, given that two RCTs in the evidence table (e_1, e_2), state that the contraceptive pill (T_1) is inferior to no treatment (T_2) for the outcome indicator breast cancer (k) the inductive arguments $(\{e_1\}, T_1 < T_2)_k$, $(\{e_2\}, T_1 < T_2)_k$, and $(\{e_1, e_2\}, T_1 < T_2)_k$ can be generated. In a nutshell, their procedure can be viewed as taking the last of these inductive arguments as being the relevant one; i.e. the inductive argument with the broadest set of evidence for the respective superiority, inferiority, or equitability claim for an outcome indicator.

With the notion of an inductive argument in place, we can specify the aggregation rule for a single outcome indicator. Consider the example of two studies with contradictory claims about the effects of contraceptive pills. Let us assume that two inductive arguments can be generated for the outcome indicator 'ovarian cancer': $(\{e_1\}, T_1 < T_2)_k$ and $(\{e_2\}, T_1 > T_2)_k$, where T_1: contraceptive pill and T_2: no treatment. Hunter and Williams propose two options to aggregate this information.[14] First, they suggest performing a statistical meta-analysis which aggregates the effect sizes of the two studies. If the effect sizes are aggregated into one, then one can read off the binary superiority relation. This amounts to generating a new inductive argument (while deleting the two other ones) with the evidence basis $\{e_3\}$ being the meta-analysis: $(\{e_3\}, T_1 < T_2)_k$. In this way, the conflicting evidence has been removed from the evidence table. Second, the reliability criterion could be contradicting one (or both) of the inductive arguments and thereby resolve the conflict between the two arguments. This strategy appears a bit hopeful. It seems that Hunter and Williams assume that all (or certainly the largest part) of the studies will give the same ordering for the treatments for any particular indicator. Note that they do not propose an alternative aggregation rule which is salient: a majority (or supermajority) rule could be used in assessing the relevant majority proportions between numbers of studies indicating that T_1 is superior (inferior) to T_2.

So far, we have described the first aggregation step. Now, we briefly turn to their *treatment of the overall multi-criteria problem*: What is the best treatment option taking into account all relevant outcome indicators (e.g. ovarian cancer, breast cancer, and pregnancy)? Hunter and Williams suggest a pairwise comparison of the inductive arguments involving different outcome indicators (e.g. the comparison of two inductive arguments; one stating that T_1 is superior to T_2 with respect to ovarian cancer, the other stating that T_2 is superior to T_1 with respect to pregnancy) (Hunter and Williams 2012, p. 15). This pairwise comparison is based on the importance of the involved outcome indicators as well as the risk of developing the outcome indicators (ibid.). This process effectively identifies the winning inductive argument, which states as its conclusion an ordering of the treatments (ibid., p. 19).

Let us briefly take stock. Hunter and Williams propose a decision support tool that has the components outlined in Sect. 2. In particular, the evidence aggregation component can be seen to accord with a witness model. So the right building blocks are present in Hunter and Williams' aggregator, but clearly there are many more design choices one might query. Indeed, Hunter and Williams' evidence aggregator

[14] This was clarified in personal communication with Hunter and Williams.

may strike the reader as rather crude;[15] evidence is delineated simply in terms of separate studies; there is a small set of meta-arguments, which form the basis for the reliability judgements; and, most importantly, the binary reliability judgments do not allow ruling in favour of one evidence piece over another in degrees. But perhaps this relatively crude structure does in fact strike the right balance between various desiderata, allowing an optimal amount of automation. In general, how should we even begin to make such an assessment? We turn to this underappreciated and yet very important issue in the next section.

4 A new criterion for determining the optimal degree of automation: capacity for robustness analysis

Like other approaches to evidence aggregation in the literature, Hunter and Williams' proposal can be seen to accord with a witness model, but this in itself does not go far in terms of settling the quality of the inferences. We now consider in more detail whether their procedure is justified or fit for purpose. We address this issue not solely for Hunter and Williams' proposal but with respect to automated evidence aggregation procedures more generally: What is the appropriate extent of automation for an evidence aggregator to help facilitate policy recommendations in, say, a medical context?

The large question that is left open by the 'witness schema' (i.e., that goes beyond the bare inferential framework), is how to delineate the independent 'witnesses', and assess the nature and reliability of 'their' findings. Reliability, for instance, is a matter of *both* the quality and relevance of the experiment or 'witness' given the hypotheses at hand. One must determine what features of individual pieces of evidence should inform this complex reliability assessment and how exactly reliability should be determined on the basis of these features. Assessing the reliability of the witnesses is one of the key things that is automated in automated evidence aggregators. This can be automated to varying extent. Recall that Hunter and Williams automate the reliability assessment via meta-arguments.

In Sect. 4.1 below we consider this question first from an ideal perspective, free from any constraints on computational resources. In effect, we consider better and worse treatments of (or inferences based on) a fixed amount of evidence (the more the better). Here we introduce our key idea of ability to perform an adequate robustness analysis. Then in Sect. 4.2 we go on to consider the more realistic scenario where there may be constraints on computational resources. We depict this as a *further* problem of a trade-off between nuanced analysis of evidence and volume of evidence. We argue that it may *not* be the case that aggregators that can handle more evidence are better. In fact, that very consideration is pivotal in assessing and comparing aggregators. We will argue that there is no prerogative to aggregate the total evidence available if this would in fact reduce overall accuracy.

[15] Admittedly, Hunter and Williams leave the question open as to whether aspects of their algorithm, in particular, the reliability assessments, should be more detailed.

4.1 Assessing aggregators in an ideal setting: no computational constraints

Absent any computational constraints, the goodness of an automated evidence aggregator is all about ideal performance—the idea is to make the best or wisest inferences possible for a class of cases (i.e., the type of evidence aggregation problem at hand), given all the available evidence. So we want the reasoning process to be as nuanced as possible, where the more of this nuance that can be captured by an explicit algorithm and so automated, the better. A higher degree of automation is desirable, all else being equal, because it allows transparency, removes computational error, enhances speed, and facilitates analysis of the sensitivity of results to choices of parameter values, i.e., robustness analysis. In the ideal setting then, any part of the reasoning process that can be made explicit in advance of seeing the particular evidence at hand, should indeed be made explicit. The only reason to leave some aspects of the reasoning process as a black box is if there are aspects of the procedure where it is less advantageous to try to specify the reasoning in advance; better to leave it to experts (assuming they are experts of average competence relative to the appropriate group) to interpret and weigh the particular evidence when it arises.

It is one thing to state the goal of automating all reasoning that can be made explicit without sacrifice in nuanced analysis, but of course it is another thing to make these extremely difficult judgments. We cannot hope to provide detailed advice on how to make such judgments in this paper. For one thing, much of the detail will depend on the type of policy task at hand and the kind of evidence available. Instead, what we seek are *strategies* that a practitioner may employ to approach or frame the question in a way that may assist, in a modest way, in arriving at an answer. To put it differently: What general criterion provides the best avenue for assessing optimal automation?

The main strategy or criterion we propose for assessing the degree of automation of an evidence aggregator can be stated as follows: Does the automated aggregator allow conducting a robustness analysis that would yield a thorough and compelling survey of the possibility space? Of course, robustness analysis has already been mentioned above as a useful by-product of automation. We initially elaborate this point (4.1.1). Our novel proposal, however, is that this consequence of automation should serve as a focal point in the algorithm design (4.1.2). This is not just a matter of wise implementation; as mentioned, we are already assuming the implementation of the algorithm in question allows parameter values to be changed by the user. It is rather a matter of settling on the explicit algorithm structure with an eye to whether the subsequent robustness analysis will serve as a reasonable survey of the possibility space for the type of aggregation problem at hand. In short, focussing on the ability to conduct an adequate robustness analysis serves to direct one's priorities to what really matters—away from the precise 'dial settings' of an aggregator, so to speak, and towards whether we have the right dials to begin with.

Before proceeding, let us first clarify a couple of terms concerning the structure of an automated evidence aggregator that will be central to our discussion in this section. First, there are *input variables* describing the evidence. Recall, for instance, Hunter and Williams' evidence table: the columns are the input variables accounting for relevant features of the evidence, and each row—an individual piece of evidence—is effectively a vector of values for these input variables.

Second, there are the *parameters* of the aggregation function that dictate how the values of the input variables bear on the assessment of each piece of evidence and ultimately on the overall aggregation or final inference concerning the hypotheses in question. For instance, the aggregation function might include a parameter 'threshold sample size' which is used to measure the quality of a piece of evidence with respect to sample size. These parameters are associated with a range of possible *values*. Thanks to programming design, the parameters can be set to any value within this range/set, depending on initial user input.

4.1.1 Robustness analysis in its traditional role: as a useful upshot of automation

As a general characterisation, robustness analysis involves determining the stability of a result given changes in underlying assumptions. Two forms of robustness analysis are widely distinguished in the literature: *Derivational (or inferential) robustness analysis* looks at the stability of model derivations (or inferences) given changes in the model (or background) assumptions (Kuorikoski et al. 2010, p. 542).[16] *Measurement robustness analysis* looks at the stability of empirical results given changes in empirical modes of determination (such as different types of experiments) (Wimsatt 1981, p. 128).

Here, we focus on derivational robustness analysis (DR) since it is a form of error analysis, i.e. a way of exploring the sensitivity of results to choices of parameter values. This role of DR is also called *heuristic function* as it allows a transparent and traceable way of dealing with unavoidable idiosyncratic choices in the construction of a model (here, evidence aggregators), due to uncertainty and/or reasonable disagreement. For DR allows determining the relative importance of various components of a model with respect to the output variable of interest (Kuorikoski et al. 2010, p. 543).

Automated evidence aggregators, given our assumption about wise implementation, are well set up for this kind of error analysis. This point has been brought up by Hunter and Williams in relation to the meta-arguments in their aggregator. They note that their procedure allows "a form of sensitivity analysis" (Hunter and Williams 2012, p. 25) by including different meta-arguments. By including different meta-arguments, in effect, the reliability of the evidence can be assessed in a range of ways, and the impact of these differing assessments on the resulting inferences can be monitored.

In other contexts too, sensitivity analysis is recommended as a way to keep track and explore the implications of model choices that are subject to uncertainty/reasonable disagreement. For instance, Stegenga (2011, p. 498) points out that there are many choices of this kind in statistical meta-analysis:

> Meta-analysis fails to constrain intersubjective assessments of hypotheses because numerous decisions must be made when performing a meta-analysis which allow wide latitude for subjective idiosyncrasies to influence the results of a meta-analysis. (ibid.)

[16] This is also frequently called sensitivity analysis (Raerinne 2013, pp. 287–288).

4.1.2 Robustness analysis in its new role: as a central design criterion

For all the good of robustness analysis, one might regard it a secondary issue when it comes to assessing the degree of automation of an evidence aggregator. Surely the primary issue is whether the aggregator facilitates roughly the best inferences possible given the available evidence; error analysis is a matter of extra detail. As suggested above, robustness is indeed typically considered an ex post analysis or a way to check what confidence one should have in a model result. Here we defend, however, a more central role for robustness analysis in the construction and assessment of an evidence aggregator. In short, the *prospect* of what robustness analysis can be performed focuses one's attention on what really matters—the functional form and possible inputs to the evidence aggregator, rather than the precise parameter values featuring in the aggregator. In other words, the prospect of robustness analysis helps one to assess what parts of the inference process can be made explicit and transparent.

There are at least two reasons why focussing on the capacity for robustness analysis is helpful in making these judgments about algorithm design: (a) it allows one to recognise that certain types of uncertainty/error (regarding precise parameter values) *do not* compromise automation, since the impact of these uncertainties can be explored via the robustness analysis, and (b) it allows one to recognise that other types of uncertainty/error *do* in fact compromise automation; cases where there is not only low confidence in the 'best-guess' estimates, but in fact there is low confidence in the entire possibility space that would be afforded by robustness analysis. In this case, it is not clear whether the remedy is more or less automation, but robustness analysis can guide the deliberation process.

Let us clarify the latter deliberation process by appeal to an example. It helps to imagine the incremental development of an aggregator. The starting structure might be a very basic one, where the pieces of evidence are described in terms of two input variables, say, 'type of study', with possible values 'randomised controlled trial (RCT)' and 'observational study', and also 'statistical significance', where possible values are simply 'yes' and 'no'. The logic of the reliability assessments might be along the following lines: Only studies that are statistically significant have positive reliability (such that they are included in the aggregation), and amongst those, the RCTs are given more weight according to a parameter *beta*, specifically, RCTs are given *beta* times the reliability weighting of observational studies. (Note that Hunter and Williams introduce a crude reliability judgment of this sort by considering meta-arguments that include/exclude evidence pieces based on whether results are statistically significant or not.) Now one might reflect on the prospective robustness analysis afforded by this aggregator. The possibility space will include inferences based on a range of values for *beta*. But this might be regarded too limited and misleading a set of possibilities.

For the above example aggregator, the possibility space afforded by robustness analysis might be deemed more adequate if the algorithm for making reliability judgments were either more or less detailed. A relatively straightforward innovation is to convert judgments that are currently implicit but need not be to explicit aspects of

the algorithm. For instance, with regard to our example, the judgments of statistical significance could be spelled out more explicitly; this would involve substituting the *p value* of the study as the input variable, and then deriving whether the study is statistically significant according to the parameter *alpha*, such that if the *p value* is less than *alpha* the study is deemed statistically significant. The corresponding robustness analysis would then produce a possibility space that includes a range of values for *alpha*, which would presumably be more thorough.[17]

So much for the low hanging fruit. The more difficult judgments concern aspects of the reasoning process where it is not clear whether more or less detail in the explicit algorithm would be better. Adding detail to the explicit algorithm is a good thing provided this is tracking genuine nuance of reasoning; the alternative scenario, however, is when extra detail in the explicit algorithm systematically distorts the reasoning process, making it more rigid in a way that is not rectified by robustness analysis. Returning to our example, a key reason why the initial robustness analysis might be deemed inadequate is that the reliability weightings depend purely on 'study type' (amongst those that are statistically significant), and yet it might be thought that this is *not* the most pertinent grouping as far as quality of evidence is concerned. One possibility is to add further dimensions to this grouping: perhaps 'sample size' and a measure of the 'relevance of experimental subject', i.e., the closeness of the experimental group to the patient class at hand, could also be included as input variables, and treated in the reliability function with reference to appropriate parameters. In this case, the robustness analysis would effectively survey the possibilities associated with changing the relative weights of these more fine-grained study groupings, which would potentially be a more adequate representation of the real space of possibilities.[18] On the other hand, it might be thought that this extra detail in the explicit algorithm would only make matters worse; that the corresponding robustness analysis would yield a possibility space that is even more misleading and would not include reasonable inferences. Perhaps the reliability assessments should be shifted entirely to implicit 'on the fly' expert reasoning rather than explicitly coded. In this case, the evidence input variable would simply be 'reliability of study'. Of course, this means one could not explore the possibility space so readily, but one might at least be happier with the 'best guess' estimate.

This example of course raises a lot of further substantial questions. But we hope the central point is clear: that it is far from obvious what aspects of an evidence aggregation process are best made explicit; focussing on the prospects for robustness analysis provides some help, however modest, in settling this question.

[17] We are not hereby endorsing statistical significance as important for determining whether a study result ought to be included in evidence aggregation. There are reasons to worry about this interpretation of statistical results. Our claim is simply that *if* this property were to play such a role, better to make the reasoning as explicit as possible, and enable the exploration of changes in the chosen level of significance.

[18] For example, we can define for the input variable 'sample size' the parameter, *gamma*, such that the reliability assessment of a study correlates not linearly to the sample size but increases in a step-wise fashion. For the input variable 'relevance of experimental subject', a parameter *delta* could be introduced in relation to the similarity measure between the experimental and patient group.

4.2 Handling computational constraints: when is greater volume of evidence better?

Now we turn to the scenario where there are constraints on (or costs associated with) computational and other resources. In practice, this is always the case, and no doubt Hunter and Williams have resource constraints in mind. To start, the on-going maintenance of the database of evidential inputs demands a lot of person hours—for identifying and recording the relevant features of each study or piece of evidence. (Moreover, it may be that only some characteristics of a study are available in the first place.) Then when it comes to running the aggregator, once the database is in hand, the algorithms for selecting, assessing the reliability and aggregating relevant pieces of evidence to arrive at a conclusion about the hypotheses in question all require processing time. In addition, there may be costs associated with the end-user interpreting the methodology/ inferential output of the aggregator (an important aspect of transparency). In short, evidence aggregation typically involves a variety of costs; most importantly with respect to person hours and computer processing time.

One might suppose that the assessment of evidence aggregators when resources are costly is a very messy business. It seems that it calls for some trade-off between inferential accuracy and efficiency. And epistemology alone cannot answer the question: To what extent should we 'cut corners' in assessing the total body of evidence so as to arrive at conclusions more quickly or with less expenditure of other resources?

We want to stress here, however, that this is not quite the right way to think about assessing the performance of an evidence aggregator that is subject to resource constraints. There is no getting around the difficult trade-off between reducing resource costs and increasing inferential accuracy.[19] But one must be careful in thinking about how to spend any given resource budget to best achieve accuracy. The above subtly presupposes that all the available evidence *mus*t be taken into account. Of course, it is natural to think that all the evidence should be considered and a way to meet the resource constraints should be found in the way the evidence is analysed, the reason being one of the core tenets of evidential logic, i.e., the view that inference should be based on *all available evidence*.[20] But that tenet refers to an ideal setting where all the available evidence is taken into account in a fully appropriate and nuanced way. There is no similar logical demand to attend to all available evidence in a practical setting where, due to resource constraints, this evidence cannot all be assessed in full detail. That would indeed be an odd requirement on an evidence aggregator—that just because some apparently relevant evidence has been tabled, it must influence the inference at hand, even if in a necessarily crude fashion.[21]

[19] Ideally such a tradeoff would be modeled as a 'value of information' decision problem and thus settled in a justified way.

[20] This is referred to as the *Principle of Total Evidence* (Carnap 1947).

[21] To clarify: Our target here is the mistaken idea that all the tabled evidence must be taken into account *in a similarly detailed way* (however much detail that may involve) and have *similar potential for influence on the inference at hand*. Of course, there is another sense in which all the tabled evidence must be taken into account—even if only a subset of the tabled evidence is examined in detail, there must be a way of selecting this subset that involves searching through *all* the tabled evidence. But in this case the tabled evidence would not all be treated similarly.

Our claim is the following: The assessment of the optimal degree of automation for an evidence aggregator under circumstances of resource constraints (once the budget has been settled) is not so different from the assessment of an aggregator that is free from resource constraints. In both cases it is performance, i.e., quality of inference regarding the hypotheses at hand, that matters, and this is best assessed by focussing on the capacity for robustness analysis. In the context of resource constraints, the 'principle of total evidence' may be better honoured by processing a subset of evidence in more detail rather than a greater amount of evidence in lesser detail. That is the underlying balancing act, in any case: Are the available resources spent in the best way possible? Should the resource budget(s) be spent on greater nuance in the description and analysis of evidence, or rather on a greater volume of evidence?

We do not deny that this adds yet a further dimension to the question of what is the optimal extent of automation for an aggregator. In the ideal case, it was simply a matter of what aspects of the reasoning process could sensibly be made explicit—a difficult enough judgment in itself. Resource constraints introduce a new complication: it may be that less-than-ideal treatment of evidence enables more volume of evidence to be processed, in the interests of accuracy. This is a further balancing act, but one that may also be informed by the capacity of an evidence aggregator for robustness analysis. Let us simply note an extreme scenario that may shed light on the more difficult non-extreme cases. The worst case, so to speak, is an evidence aggregator for which we are not confident that *any* of the input evidence contributes to higher quality inference about the hypotheses at hand, regardless of the precise values of key parameters. Here it is not even the case that, conditional on using the aggregator at hand, inferences based on more evidence (requiring more resources) are better. In the more ordinary and difficult cases, by contrast, all the aggregators under assessment will be ones for which more evidence (requiring more resources) permits better inferences using *that* aggregator. (Here the possibility space associated with each aggregator, owing to robustness analysis, is at least deemed adequate.) The further question in this case is which aggregator allows for the best quality inferences given the resource budget at hand, where some aggregators process less evidence with greater nuance while others process more evidence with lesser nuance.

5 Potential objections to our criterion

We consider here potential criticisms of our approach to determining the optimal extent of automation in cases without and with resource constraints. On the one side of the spectrum, it might be objected that we rely too much on intuitive reasoning or expert judgement, when it would be better to appeal to objective assessments of the track record of evidence aggregators. On the opposite side, it might be objected that there is no plausible alternative to implicit expert reasoning when it comes to aggregating diverse evidence.

The main thing to be said in response to the first position is that it is not obvious what an objective assessment of the track record of an aggregator consists in. A clear criterion for an aggregator having produced a verified result would be an instance in which the aggregator predicts an event and the event in fact takes place. However, the

kinds of hypotheses considered in this paper (e.g. 'treatment 1 is better than treatment 2 with respect to breast cancer rates') are generally not of this nature. Medical hypotheses tend to be stochastic (and moreover, associated predictions often depend on a number of auxiliary conditions); thus there is no definitive point in the future at which evidence is received that settles such matters.

The second position might go like this: The automation of evidence amalgamation is doomed to fail because it is inherently resistant to explicit algorithms. To start, one might argue that judgment is required on a very basic level, namely for the application of any (reliability) criterion to a particular experimental finding. Even if the reliability criterion is expressed in a functional form in terms of input variables and parameters, this objection goes, the value assigned to the input variables in any concrete case requires expert judgement. Stegenga (2014, p. 203) cites empirical studies showing that applying the *same* quality assessment tools (QATs) to the *same* medical finding leads to widely diverging quality assessment of this medical evidence, presumably because the finding itself is perceived differently by different investigators. By way of response, we accept the importance of expert judgment in the process of evidence aggregation. Yet we want to stress that there is no reason to deny that at least some of the reasoning involved in evidence aggregation/inference can be broken down and made explicit via an algorithm. Indeed, our question in this paper concerns the optimal *extent* of automation.

One might otherwise argue that evidence amalgamation is resistant to explicit algorithms since every case of amalgamation is distinct, and, hence, an algorithm, which is necessarily more general, misses the distinct features of the particular case. We do not find this objection convincing. Surely, there is a sufficient degree of similarity between classes of evidence aggregation problems that allow formulating explicit aggregation rules that are broadly applicable. The QATs mentioned by Stegenga are a case in point.

6 Concluding remarks

This paper addressed the question of what is the optimal degree of automation for a given type of evidence aggregation problem. We looked at this question in the context of medical policy-making which involves large and rather diverse sets of evidence. We argued that general inferential logic, such as the witness schema, helps to frame the evidence aggregation task but does not in itself settle the further substantial question about the optimal degree of automation. We illustrated this point with the help of our case study—Hunter and Williams' highly automated aggregator.

We hope that our discussion has highlighted the difficulties in answering this question. Designing a large-scale evidence aggregator is non-trivial, to say the least, and this paper leaves a host of important questions unanswered. Of course, much will depend on the details of the type of policy task at hand; our modest aim in this paper was to provide a general strategy/evaluation criterion that could provide a focal point for making the requisite assessments. We have proposed the ability to perform an adequate robustness analysis (which depends on the nature of the input variables and

parameters of the aggregator) as the focal criterion, primarily because it directs efforts to what is most important, namely, the structure of the algorithm and thus the appropriate extent of automation. Moreover, where there are resource constraints on the aggregation process, one must also consider what balance between volume of evidence and accuracy in the treatment of individual evidence best facilitates inference. Again, concentrating on robustness analysis helps here, but there are further trade-offs between nuanced analysis of evidence and volume of evidence that must be taken into account. There is no prerogative to aggregate the total evidence available if this would in fact reduce overall accuracy. This trade-off between nuanced analysis of evidence and volume of evidence is in our view a promising line of further research.

Acknowledgements We would like to thank three anonymous reviewers for providing very detailed and constructive feedback and Anthony Hunter and Matthew Williams for fruitful email exchanges regarding their framework. Audiences at the London School of Economics as well as the CLMP 2015 in Helsinki supplied us with helpful comments on earlier versions of this paper.

Compliance with ethical standard

Conflict of interest The authors declare that they have no conflict of interest.

References

Ben-Gal, I. (2007). Bayesian networks. In F. Ruggeri, F. Faltin, & R. Kenett (Eds.), *Encyclopedia of statistics in quality and reliability* (pp. 1–6). London: Wiley and Sons.

Carnap, R. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research*, *8*, 133–148.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

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

Hunter, A., & Williams, M. (2010). Using clinical preferences in argumentation about evidence from clinical trials. In *Proceedings of the 1st ACM International Health Information Symposium* 11/2010, pp. 118–127.

Hunter, A., & Williams, M. (2012). Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, *56*(3), 173–190. [1-36].

Hunter, A., & Williams, M. (2013). Aggregating evidence about the positive and negative effects of treatments using a computational model of argument. In *LSE Choice Group Talk*, pp. 1–38, October 21, 2013.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge: Cambridge University Press.

Kuorikoski, J., Lethinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *The British Journal for Philosophy of Science*, *61*, 541–567.

Krimsky, S. (2005). The weight of scientific evidence in policy and law. *American Journal of Public Health*, *95*(1), 129–136.

List, C. (2012). The theory of judgment aggregation: An introductory review. *Synthese*, *187*(1), 179–207.

Montuschi, E. (2009). Questions of evidence in evidence-based policy. *Axiomathes*, *19*, 425–439.

NICE. (2006). The guidelines manual. London: National Institute for Health and Clinical Excellence. http://www.nice.org.uk. Accessed July 20, 2015.

Raerinne, J. (2013). Robustness and sensitivity in biological models. *Philosophical Studies*, *166*, 285–303.
Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, *21*(2), 157–170.
Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*, 497–507.
Stegenga, J. (2014). Herding QATs: Quality assessment tools for evidence in medicine. In P. Hunemann, G. Lambert, & M. Silberstein (Eds.), *Classification, disease and evidence: New essays in the philosophy of medicine* (pp. 193–211). London: Springer.
Weed, D. L. (2005). Weight of evidence: A review of concept and methods. *Risk Analysis*, *25*(6), 1545–1557.
Wimsatt, W. C. (1981). Robustness, reliability and overdetermination. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–163). San Francisco: Jossey-Bass.
Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*, 219–240.