⋆

The feasibility and malleability of EBM+

⋆

Jon Williamson

April 16, 2021
Theoria 36(2), doi 10.1387/theoria.21244.

Abstract

The EBM+ programme is an attempt to improve the way in which present-day evidence-based medicine (EBM) assesses causal claims: according to EBM+, mechanistic studies should be scrutinised alongside association studies. This paper addresses two worries about EBM+: (i) that it is not feasible in practice, and (ii) that it is too malleable, i.e., its results depend on subjective choices that need to be made in order to implement the procedure. Several responses to these two worries are considered and evaluated. The paper also discusses the question of whether we should have confidence in medical interventions, in the light of Stegenga's arguments for medical nihilism.

Keywords: Causality; Causation; EBM; EBM+; Russo-Williamson Thesis; RWT; Evidential pluralism; Medical nihilism.

§1 EBM and EBM+

Causal claims are central to medicine. All areas of medicine seek to establish such claims: in basic medical science, claims about disease progression and the maintenance of health; in exposure assessment, claims about the effects of exposure to chemicals or other agents; in intervention assessment, claims about the effects of health interventions, for example. The idea underlying evidence-based medicine (EBM) is to make the evidence for these causal claims explicit, and to make methods for evaluating that evidence explicit, in order to improve the reliability of the assessment of causal claims:

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence (Sackett et al., 1996, p. 71).

However, the way EBM seeks to achieve this goal is by focussing on clinical studies—particularly randomised controlled studies (RCTs)—and by excluding other kinds of evidence or by viewing it as inherently low quality:

> Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. (Guyatt et al., 1992, p. 2420)

The exclusion of 'pathophysiologic rationale'—more generally, of mechanistic evidence—conflicts with a recent line of research on the epistemology of causality. Russo and Williamson (2007) argued that to establish a causal claim in the health sciences one should look for evidence of mechanisms as well as evidence of correlation. This is because there are many possible explanations of an observed correlation between variables A and B: one such explanation is that A is a cause of B, but others are reverse causation, confounding, chance, or other relationships between A and B, such as semantic, constitutive, logical, physical and mathematical relationships (Williamson, 2019a, §1.2). What is distinctive about the former, causal, explanation is that there is some mechanism complex linking A to B by which instances of A explain instances of B and which gives rise to the observed correlation. Hence evidence of mechanisms is crucial to establishing causality.

Fig. 1 provides a visual representation of this epistemology of causality. Association studies are studies which test for an association between A and B; these include both experimental and observational studies and encompass studies in clinical medicine as well as epidemiological studies of disease, and systematic reviews and meta-analyses of such studies. Association studies usually test whether A and B are probabilistically dependent, conditional on other potential causes of B. Such studies provide direct evidence of a correlation via the confirmatory channel C_1. Within the class of association studies, RCTs are prized by proponents of present-day EBM because they can reduce the risk of confounding by unforeseen causes of B, so they can provide some indirect evidence of the existence of a mechanism of action (C_2). A more direct way of confirming the presence of a mechanism of action is by confirming specific mechanism hypotheses, which posit features of a possible mechanism complex linking A and B (M_2). Mechanistic studies test these hypotheses (M_1). In some cases, established specific mechanism hypotheses can also confirm or undermine the presence of a posited correlation (M_3)—see Williamson (2019a, §2.2) on this point.

This approach, then, motivates the systematic evaluation of mechanistic studies alongside association studies when assessing a causal claim in medicine. Parkkinen et al. (2018) provide a set of general procedures for performing this evaluation and call this approach 'EBM+'. Although EBM+ very much fits the spirit of EBM, because it seeks to make evidence and its evaluation explicit and systematic, it flies in the face of the actual practice of present-day EBM, which, as we have seen, devalues mechanistic studies. Parkkinen et al. (2018) view present-day EBM as a first approximation to correct evidence evaluation, with the scrutiny of mechanistic studies a further step along the path—hence the '+' in 'EBM+'. Note that EBM+ is intended to be applicable throughout medicine, which is broadly construed to include the health sciences as well as clinical practice, because it is based on a general thesis about how to establish causal claims. While Fig. 1 depicts evidential relations when assessing causation in a target population, the EBM+ programme also has a set of procedures for assessing whether claims based on studies carried out on a different source population can be extrapolated to the target population (Parkkinen et al., 2018).

The above view of the epistemology of causality has been the object of some controversy in the literature (see Williamson, 2019a, §1), and there are many who continue to agree with Guyatt et al. (1992) that mechanistic evidence should be 'de-emphasized' or ignored. For example, The Oxford Centre for Evidence-Based Medicine still (as of 2020) places mechanism-based reasoning at the bottom of its evidence hierarchy (OCEBM Levels of Evidence Working Group, 2011), and Miriam

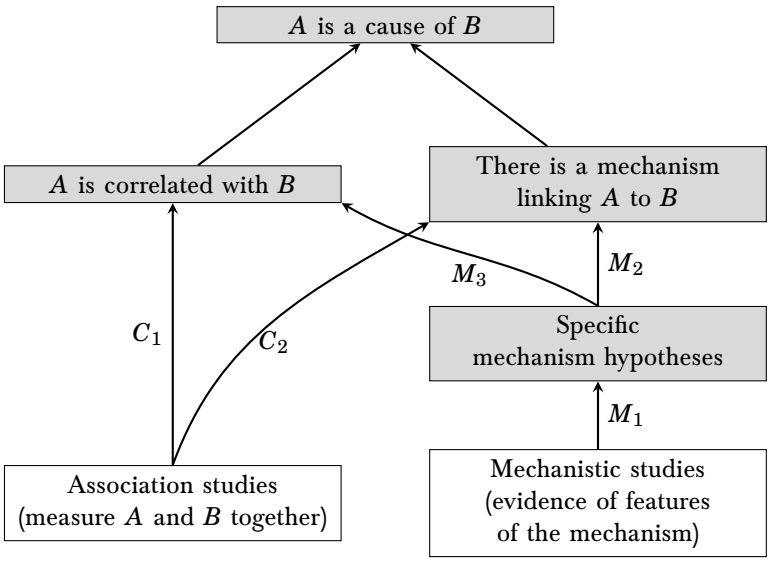

Figure 1: Evidential relationships for establishing a causal claim (Williamson, 2018).

Solomon holds that

> Mechanistic reasoning (or "mechanistic evidence") does not play a role in the process of evaluating the effectiveness of new interventions (Solomon, 2015, p. 132).

We will not revisit the rationale behind EBM+ in this paper, as it has been widely discussed elsewhere.[1] Instead we will focus on two new challenges for EBM+. One worry is that the systematic consideration of mechanistic studies may simply not be feasible—a worry that we consider in §2. Another concern is that EBM+ may be prone to manipulation by subjective influences, which we consider in §3. Finally, in §4, we discuss some consequences of our findings for Jacob Stegenga's claim that one should have little confidence in the effectiveness of medical interventions.

§2 The feasibility of EBM+

One question that immediately faces the EBM+ programme is its feasibility. It is hard enough to systematically assess association studies, which are well indexed in databases and which are amenable to standardised statistical analysis. Mechanistic studies, however, are very heterogeneous and are not in general well indexed. Is it really practical to systematically evaluate mechanistic studies alongside association studies?[2]

[1] See Williamson (2019a, §1) for references. The EBM+ programme is very much in line with the efforts of Cartwright and Hardie (2012) to improve evidence evaluation in evidence-based policy.

[2] Howick (2011, §10.4), for example, expresses doubts about feasibility on the grounds that mechanistic evidence is usually too incomplete, or the mechanisms themselves too complex, to be able to usefully consider mechanistic evidence. La Caze (2019, §3.2) also presents a feasibility-related challenge, namely that of spelling out how evidence of complex mechanisms can inform extrapolation inferences.

The obvious way to meet this feasibility challenge is to provide a good example of evidence assessment in medicine that appeals to EBM+ or something like it and that is clearly feasible. In this section we shall investigate whether such an example is to be found and, if so, where it is to be found.

Intervention assessment. It turns out that there are very few examples of the systematic and explicit evaluation of mechanistic evidence when assessing claims about the effectiveness of interventions.

One potential example is the umbrella review (i.e., review of systematic reviews and meta-analyses) of Posadzki et al. (2018), who assess effects of melatonin on health. They consider mechanistic evidence alongside association studies and formulate and evaluate specific mechanism hypotheses. However, their review of mechanism hypotheses is limited to the goal of identifying potential mechanisms of action; they do not integrate the conclusions of their analysis of the mechanism hypotheses with the results of their analysis of association studies in order to come up with an assessment of the causal claim based on all the evidence. Thus this study cannot be said to implement the EBM+ approach in its entirety. Nevertheless, they show that it is feasible to perform a systematic review in order to identify a range of specific mechanism hypotheses, which is an important component of the EBM+ programme.

Another potential example is the assessment of a pegylated combination therapy of peginterferon alfa and ribavirin for the treatment of chronic Hepatitis C, which led to its recommendation as the optimal treatment. Auker-Howlett and Wilde (2019) show that the reasoning that justified this recommendation can only be understood by means of the conceptual apparatus of Fig. 1. This is because neither association studies on their own nor mechanistic studies on their own provided grounds for the recommendation: only when association and mechanistic studies are considered in combination with one another is the recommendation warranted. While their argument is compelling, it shows only that one needs the conceptual apparatus of EBM+ in order to account for intervention assessment here—it does not show that the detailed procedural recommendations of EBM+ are feasible. This is because the UK National Institute for Health and Care Excellence (NICE), who issued the recommendation about the treatment, cited only association studies in support of their recommendation (NICE, 2010), in accordance with EBM procedure but not with EBM+ procedure.[3] (Given the arguments of Auker-Howlett and Wilde (2019), the association studies on their own should not have been taken to establish effectiveness, and EBM+ procedure would have recommended a review of mechanistic studies on this occasion.) Thus, we must search elsewhere for evidence of the feasibility of the full EBM+ programme.

Disease assessment. Mechanistic evidence is routinely considered in disease assessment: there is often some integration of mechanistic considerations with the assessment of epidemiological studies in order to obtain an overall assessment of a claim about disease causation. However, this integration can be rather haphazard. To give an example, reviews that assessed whether Zika virus causes birth defects considered mechanistic evidence in several different ways (Williamson, 2018). Frank et al. (2016) used the well-known Hill indicators of causality to assess teratogenicity (see Table 1): these indicators include 'Plausibility' and 'Coherence' which assess

[3]However, there may have been some undisclosed mechanistic reasoning. See §4 on this point.

Table 1: The causal indicators of Hill (1965).

Strength	Strength of the observed association
Consistency	Consistency of the observed association
Specificity	A narrowly defined cause and effect (disease), and the cause is not associated with other diseases
Temporality	The putative cause occurs before early stages of the disease
Biological gradient	A dose-response curve
Plausibility	Plausible given the biological knowledge of the day
Coherence	No conflict with the known history and biology of the disease
Experiment	Confirming experimental evidence
Analogy	Similar effects of similar causes

fit of the causal claim to mechanistic considerations. Rasmussen et al. (2016), on the other hand, used Shepard's indicators, which are tailored specifically to the assessment of teratogenicity and which include a single indicator that considers fit to established mechanisms. Meanwhile Krauer et al. (2017) used an ad-hoc set of indicators, which included biological plausibility. None of these systems specifies exactly how the various indicators combine and different reviews came to different conclusions about teratogenicity. In sum, then, while these reviews do support the claim that the routine assessment of mechanistic consideration is feasible, they are far from exemplars of the full EBM+ programme, which seeks a more systematic integration of evidence.

Exposure assessment. The assessment of the effects of exposures provides a more fruitful hunting ground for evidence of the feasibility of EBM+. In particular, the new (2019) methods of the International Agency for Research on Cancer (IARC) for evaluating the carcinogenicity of various agents, presented in IARC (2019a) and Samet et al. (2020), are very much in line with EBM+. Fig. 2 provides an EBM+-style conceptualisation of the evidential relationships in IARC's procedure. IARC separately assesses exposure studies, human studies, animal studies and mechanistic studies and then integrates these assessments to come up with an overall evaluation of carcinogenicity. Human studies encompass epidemiological studies on humans; these provide direct evidence of correlation and indirect evidence of the existence of a mechanism of action, as portrayed in Fig. 1. Similarly, mechanistic studies are used to directly test specific mechanism hypotheses: IARC have developed a list of ten key characteristics of carcinogens and mechanistic studies are used to determine which of these characteristics, if any, is present in mechanisms involving the agent in question. Additionally, animal studies provide indirect evidence of both correlation and mechanism in humans, as long as the relevant animal and human mechanisms are sufficiently similar—a judgement that is informed by mechanistic studies (Wilde and Parkkinen, 2019).

As a recent example of the procedural aspects of IARC evaluations, consider the evaluation of the carcinogenicity of three chemicals, namely styrene, styrene-7-8-oxide and quinolene, which took place on 20–27 March 2018 and which culminated in the monograph IARC (2019b). Including those responsible for the evaluation,

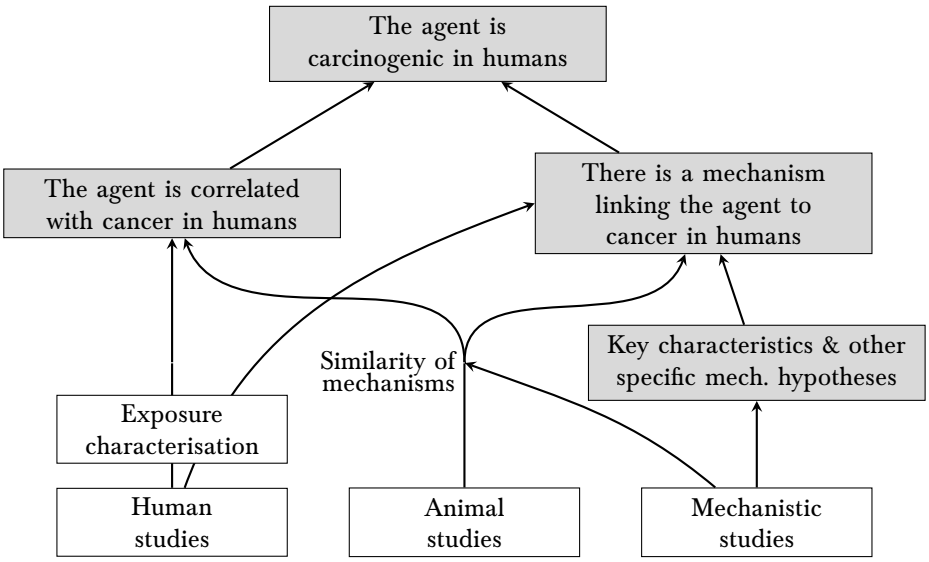

Figure 2: Evidential relationships for IARC's evaluations (Williamson, 2019b).

members of IARC who assisted the evaluation, and invited specialists, the subgroup charged with analysing exposure data had 6 members, the subgroup assessing human studies had 10 members, that assessing animal studies had 4 members, and that assessing mechanistic studies had 15 members. The size of each subgroup reflects the range of expertise required for that component of the evaluation. Before the evaluation meeting, roughly a year's preparation went into organising the meeting, selecting the studies for review, and producing an initial review of the material by the working group members. At the review meeting itself, the subgroups met separately for the first four days in order to assess the evidence in their category for each of the three chemicals. All subgroups then came together to integrate the individual assessments and generate an overall evaluation of the carcinogenicity of each of the three chemicals—this phase of the procedure took approximately three days. After the meeting, participants worked for over a year to finalise the resulting publication of the assessment, IARC (2019b). This gives an indication of the scale of the enterprise. That it is feasible and reliable is witnessed by the fact that IARC has conducted over a thousand evaluations to date, and that IARC evaluations are relied on around the world to influence public health policies that restrict exposure to carcinogens.

That IARC evaluations are broadly in line with EBM+ procedure can also be seen with the aid of Fig. 3 and Fig. 4, which characterise the EBM+ approach to evaluating efficacy and external validity respectively. In the context of an IARC evaluation, the efficacy question asks whether the agent under review is a cause of cancer in humans, while the external validity question asks whether the conclusions from animal studies extrapolate to humans. Consider Fig. 3 first. If human studies suffice to establish or rule out carcinogenicity, then, according to IARC procedure, carcinogenicity is decided and the results of the assessment of mechanistic studies do not bear on the overall evaluation. Otherwise, the IARC classification of car-

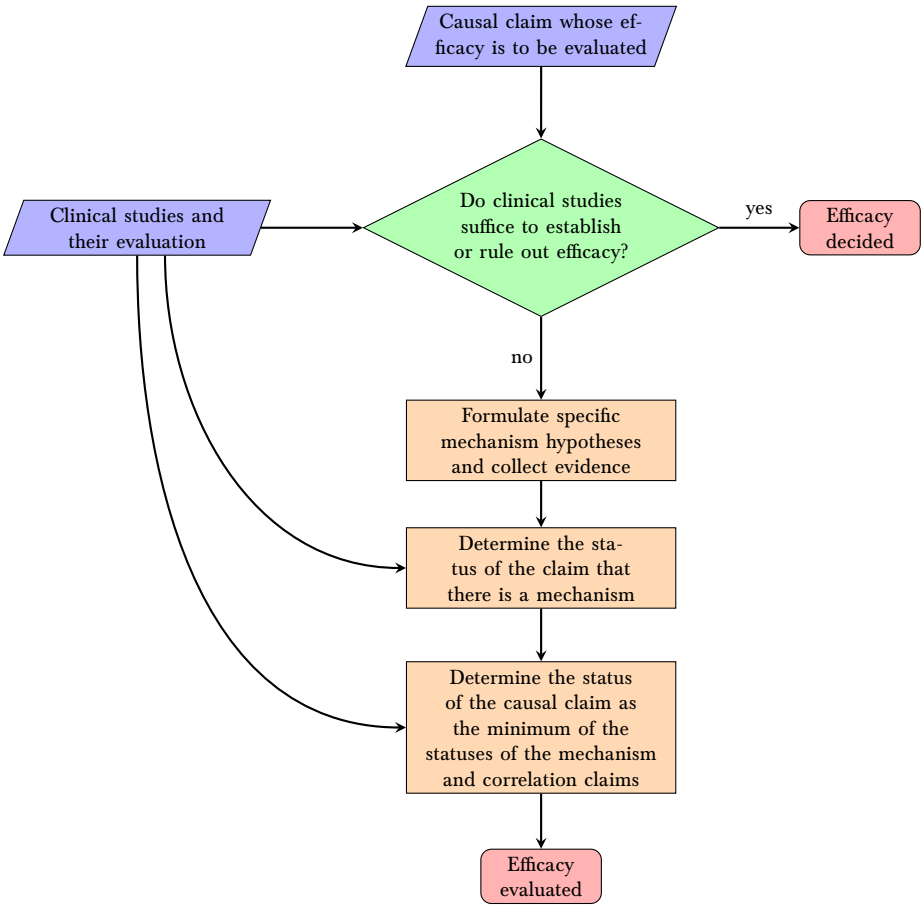

Figure 3: EBM+ procedure for assessing efficacy (Parkkinen et al., 2018, §3.3).

cinogenicity is influenced by their assessment of specific mechanistic hypotheses—the 10 key characteristics of carcinogens—and their systemic review of mechanistic studies that are relevant to these hypotheses. The subgroup responsible for assessing mechanistic studies then assesses whether these characteristics are present and whether there is strong mechanistic evidence overall, i.e., whether there is strong evidence arising from mechanistic studies for the claim that there is a mechanism of action. Thus far, IARC procedure is perfectly in line with EBM+. EBM+ then suggests that one should explicitly consider whether human and animal studies provide strong indirect evidence that there is a mechanism of action, in order to determine the status of the general mechanistic claim, and that the status of the overall carcinogenicity claim tracks the status of whichever of the correlation and mechanism claims is weaker. It is here that IARC procedure departs slightly from that of EBM+ (Parkkinen et al., 2018, Chapter 8): IARC's method for determining the overall evaluation is rather more intricate, and, according to Williamson (2019b), has certain limitations. Despite these minor differences, IARC procedure is very close to that of EBM+. Fig. 4 tells a similar story: for IARC, the role of animal studies depends on how decisive they are in determining carcinogenicity in

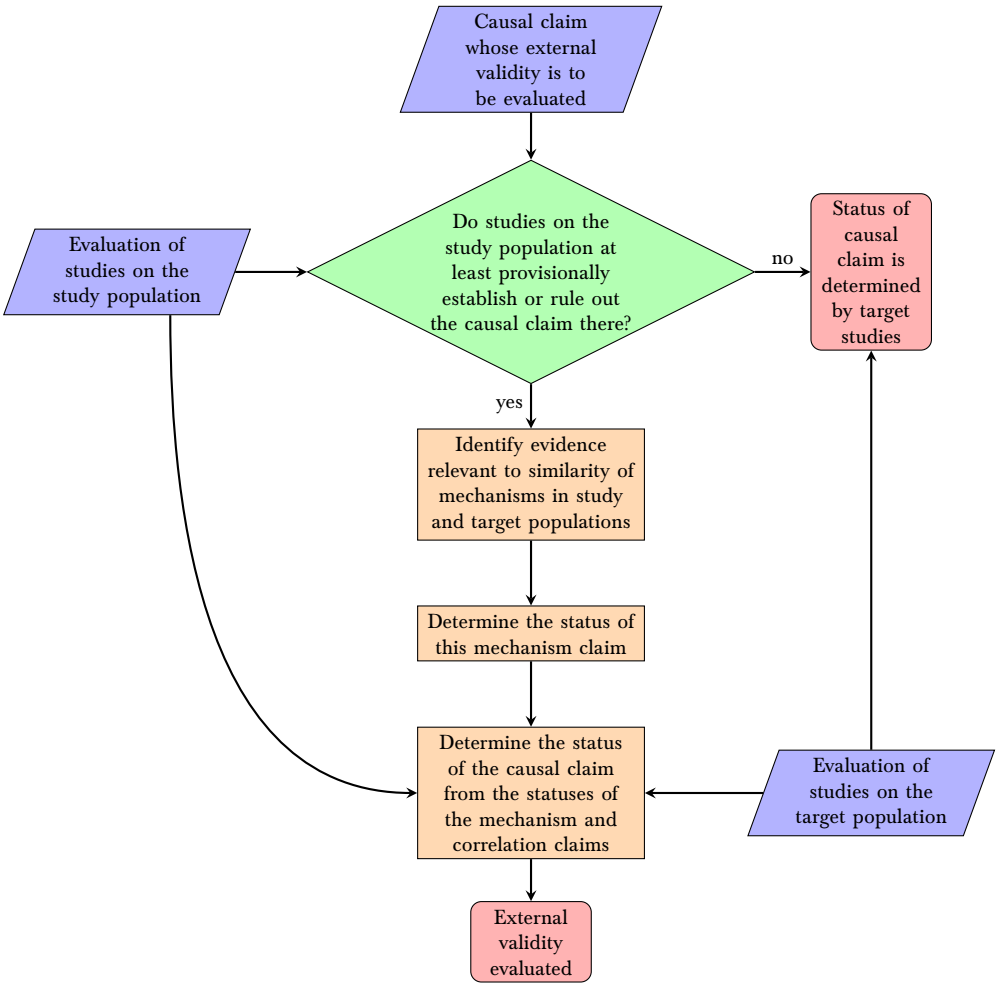

Figure 4: EBM+ procedure for assessing external validity (Parkkinen et al., 2018, §3.3).

the experimental animals and how similar the putative mechanisms of action are in animals and humans, which is in accord with the EBM+ approach of Fig. 4.

Overall, then, the feasibility of IARC procedure supports the feasibility of the EBM+ approach. IARC clearly show that it is possible to search for and assess mechanistic studies in a systematic way, and to integrate this assessment with those of epidemiological studies in humans and animal studies to determine an overall evaluation. All this requires effort: a working group tasked with assessing mechanistic evidence. But this effort is proportionate to that expended on the assessment of human and animal studies. Furthermore, EBM+ procedure imposes a lower burden than IARC procedure, because EBM+ recommends a full evaluation of mechanistic studies only where association studies on their own fail to establish causation (see Fig. 3 and Fig. 4), while IARC evaluate mechanistic studies even in situations where this evaluation cannot influence the overall assessment of carcinogenicity.

To summarise, evidence of the feasibility of EBM+ is hard to find in the areas of intervention assessment and disease assessment, but there is good evidence arising from IARC practice in exposure assessment. Although evidence of the feasibility of EBM+ is strongest in the area of exposure assessment, the domain of application of EBM+ is not restricted to exposure assessment. EBM+ offers a general methodology for assessing causation in medicine, and its feasibility for exposure assessment supports its feasibility for intervention assessment, disease assessment, and indeed basic medical science.

While the number and quality of relevant mechanistic studies will vary from area to area, IARC practice shows that, in an area where there are often very many relevant mechanistic studies, it is feasible to search for and assess these studies, and to integrate that assessment with assessments of other studies in order to determine the overall status of a causal claim of interest.

That EBM+ is feasible in exposure assessment carries directly over to disease assessment. Of course, diseases caused by infectious agents rather than chemical exposures require considering studies of mechanisms of infection and of the body's defences against infection, rather than studies of metabolism of chemicals and the effects of resulting metabolites. However, IARC already routinely considers infectious causes of cancer, and no new feasibility concerns arise there.

Let us turn next to intervention assessment. Intervention assessment differs from disease assessment insofar as experimental studies become more practical when assessing the effects of an intervention. This complicates the evaluation of association studies. However, it does not significantly complicate the evaluation of mechanistic studies. One will often need to consider mechanisms of compliance with the intervention in addition to the mechanism of action of the intervention and any counteracting mechanisms. However, mechanisms of compliance are analogous to mechanisms of exposure in exposure assessment, and these are routinely considered in some detail by IARC, for example, so no new concerns about feasibility emerge.

Thus the feasibility of IARC assessments supports the feasibility of EBM+ in both disease and intervention assessment, in addition to exposure assessment.

§3
The malleability of EBM+

Stegenga (2018) argues forcefully that the methods of present-day EBM are malleable, in the sense that their implementation requires many subjective choices and this makes their results prone to influence by interested parties. The worry arises that the methods of EBM+, which require the assessment of mechanistic studies in addition to the assessment of association studies mandated by EBM, will be even more malleable. If so, its malleability would be a serious objection to EBM+.[4]

Stegenga argues that the malleability of present-day EBM stems from a range of problems with association studies which leave them open to bias and fraud. The aim of this section is not to defend EBM from these criticisms, but to investigate whether EBM+ is more or less prone to malleability than EBM. Our approach will

[4] Howick (2019, p. 178), for example, expresses concerns about malleability. He criticises EBM+ for not focussing on the problem of financial biases, saying, 'by ignoring the problem, they cannot possibly solve it.' The extensive literature on values in science reinforces concerns about malleability—see Douglas (2009); Teira and Reiss (2013); Andreoletti and Teira (2019) and Holman (2019) for example. Gillies (2019a) responds to Holman's concerns about EBM+.

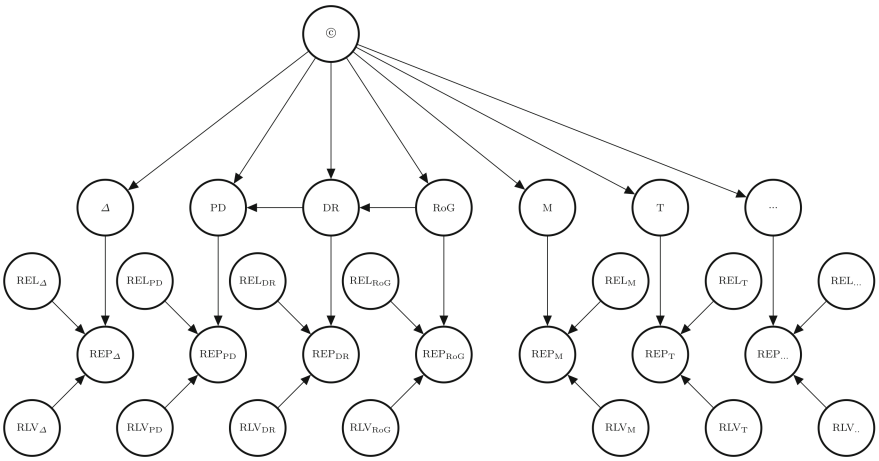

Figure 5: The Bayesian network formalisation of Hill's indicators of Landes et al. (2018).

again to be to consider various possible responses to this objection, in order to determine whether EBM+ can offer a viable defence.

Formalisation. One might attempt to respond to the malleability worry by developing a formal framework for EBM+, in the hope that formalisation reduces the scope for subjective influence. That this is not a promising strategy, however, can be seen from efforts to formalise Hill's indicators of causality, depicted in Table 1. As we saw above, Hill's approach takes mechanistic considerations into account, but Hill does not specify exactly how to tell when an indicator is present, nor how to integrate evidence arising from multiple indicators (Hill, 1965), leaving his approach open to the charge of malleability. There have been some interesting attempts to formalise Hill's approach in order to reduce malleability.

One line of work here is the *E-synthesis* approach of Landes et al. (2018), which formalises Hill's indicators by means of the Bayesian network framework. Fig. 5 represents the graph of such a network. The node at the top refers to the causal hypothesis. At the next level down there are variables related to Hill's indicators: Δ refers to difference making, PD to probabilistic dependence, DR to dose-response relationship, RoG to rate of growth, M to mechanisms, T to temporality, and so on. Each indicator is connected to an evidence report (REP) variable for every item of evidence that bears on that indicator (although only one such variable for each indicator is depicted in Fig. 5). Each evidence report variable is in turn connected to two further variables denoting the relevance, RLV, and reliability, REL, of that evidence report. A Bayesian network also requires the probability distribution of each variable conditional on its parent variables in the graph. These, according to Landes et al. (2018, p. 33), should be based on domain knowledge elicited from experts. It is here that the malleability problem emerges. Domain knowledge is very unlikely to fully determine all these probability parameters, and decisions will need to be made as to how to fill in the gaps. Bias and fraud can enter the picture here, and will be all the harder to detect because of the complexity of the formal

framework. One might attempt to elicit the probability distributions themselves from experts, but this is a big ask. Moreover, different experts are likely to provide very different probability distributions, because experts have radically differing views about the relative importance of Hill's indicators, and because many variables denote epistemological qualities—reliability and relevance—which domain experts are not used to quantifying. Hence, regardless of the merits of the E-synthesis approach, there remains plenty of scope for subjective influence and this approach is unlikely to help with the specific problem of malleability.

Swaen and van Amelsvoort (2009) also attempt to formalise Hill's indicators in order to reduce the influence of subjectivity when deciding how the various indicators should be weighed. They appeal to a 'weight-of-evidence' approach (Weed, 2005), which requires quantifying the extent to which each of Hill's indicators is met and quantifying the relative importance of each indicator. Instead of eliciting these weights from experts, which would be susceptible to the objections offered above, Swaen and van Amelsvoort (2009) try to learn the parameters of the model from a dataset consisting of past IARC evaluations of the carcinogenicity of various exposures. The problem with this approach is that their assessment of the extent to which past IARC classifications exhibit each of Hill's indicators is rather arbitrary—certainly there is room for subjective disagreement there. Moreover, it is not clear how the resulting model can be reliably extrapolated to future IARC classifications, let alone to the assessment of interventions. This is because certain model assumptions appear very questionable: for example, it is assumed that weights of indicators combine linearly and do not vary from context to context.

Thus formalisation offers little scope for ameliorating malleability. Whether one attempts to elicit model parameters from experts or to learn them from data, subjective judgements play an important role. Introducing strong model assumptions offers one way of reducing the number of free parameters that are open to subjective influence. However, this tactic merely threatens to trade malleability for unreliability, and in any case, model assumptions are themselves open to disagreement.

Quantity of evidence. A more promising response to the malleability objection notes that increasing the *quantity* of evidence tends to reduce subjective influence. The general idea is that if evidence E makes some set C_E of conclusions rationally permissible, where different subjective choices along the way can lead to different conclusions, then additional evidence F is likely to lead to a smaller, rather than larger, set C_{EF} of rationally permissible conclusions. In our context, the conclusions relate to a proposition of the form A *is a cause of* B. If we take E to be evidence from association studies and F evidence from mechanistic studies, this phenomenon suggests that the influence of subjectivity is likely to be reduced, rather than increased, by taking mechanistic studies into account.

In a Bayesian framework, where C_E might represent a set of rationally permissible degrees of belief in the proposition that A is a cause of B, this phenomenon is called the 'washing-out of priors' and is made precise by means of a range of convergence-of-opinion theorems (see, e.g., Dorling, 1975, §11; Dorling and Edgington, 1976). Under certain conditions, one can guarantee that a set of rational degrees of belief will converge to a single rational degree of belief as the quantity of evidence increases. However, these conditions are somewhat idealistic and there is room for debate about the extent to which they are met in practice (Earman, 1992, Chapter 6). Moreover, even if they are met, they provide no guarantee that

the influence of subjectivity will reduce in the short term. Indeed, there are many intriguing cases of 'dilation', where learning something new can enlarge the set of permissible degrees of belief.[5]

Regardless of these challenges, there must be something of substance to the general phenomenon, for otherwise there would be no advantage to gathering more evidence. The upshot is that one should expect EBM+ to reduce, rather than increase, malleability, in comparison to EBM. To suggest otherwise would require some reason for thinking that mechanistic and association studies together are particularly likely to lead to dilation. However, the opposite appears to be the case, as we shall now see.

Variety of evidence. An even more promising response to the malleability objection appeals to the potential for *diverse* evidence to reduce subjective influence.[6] As noted above, causal inference is beset by the problem that an observed correlation has a wide range of potential explanations, including bias and confounding, and that one can only establish causation where these other potential explanations can be ruled out. Association studies can provide some evidence against these alternative explanations. But, as Stegenga argues, association studies are prone to error and bias and are malleable. The standard view is that where the existing association studies are inconclusive, more association studies are called for. However, further association studies are prone to the *same* kinds of errors, biases and malleability as the original association studies. Just as independent witnesses are given more weight than witnesses with common interests and similar flaws, studies with different designs and carried out by teams with different interests would be more helpful.

Mechanistic studies are just such studies. Mechanistic studies tend to be much more heterogeneous than association studies and typically involve methods other than those used by association studies, such as in vitro lab work, biomedical imaging, autopsies, animal experiments and simulations. Defects of these different kinds of study are independent of defects of association studies and of each other. They are often conducted by research teams with different interests to those who carry out association studies (which tend to be carried out by drug companies seeking approval for lucrative new drugs). Thus publication bias, fraud, and industry manipulation are less of a concern for mechanistic studies than for association studies. To be sure, the teams carrying out mechanistic studies do have interests, but these interests tend to differ.[7]

To some extent, then, association studies and mechanistic studies act as independent witnesses—certainly more so than do association studies and yet more association studies. Scrutinising and evaluating both kinds of study can only help to diminish the scope of malleability and error.

Reinforcing evidence. While the quantity- and variety-of-evidence responses go some way towards addressing concerns about malleability, more can be said. There is an important sense in which association studies and mechanistic studies have

[5] See Zhang et al. (2018) for a recent discussion of the relation between dilation and disagreement.

[6] Again, there are Bayesian explications of this phenomenon (e.g., Landes, 2020), but again, there are exceptions to the general phenomenon.

[7] This is of course not to say that interests never coincide, nor that bias, fraud and industry manipulation are never a concern for mechanistic studies. See Fugh-Berman (2013); Green (2015) and Conradi and Joffe (2017) on this point.

complementary strengths. As we have observed, association studies on their own can be unreliable indicators of causality because of biases, unforeseen confounding etc. Mechanistic studies help to address precisely these deficiencies: they tell us about potential confounders and help to determine whether a correlation is genuinely causal. On the other hand, mechanistic studies on their own can be unreliable indicators of causality in two ways: (i) it can be hard to determine from a complex mechanism whether the putative cause actually makes a difference to the putative effect (the 'problem of complexity'); (ii) there may be unforeseen counteracting mechanisms which cancel out the influence of some positive mechanism of action (the 'problem of masking'). Association studies ameliorate both these problems: (i) they can be used to demonstrate the existence of a net association across the mechanism as a whole, showing that the cause does make a difference to the effect; (ii) a positive association provides evidence that unforeseen counteracting mechanisms do not fully cancel out the mechanism of action.

Thus association studies and mechanistic studies are not fully independent witnesses: they are better than independent witnesses, because they make up for one another's deficiencies. From an epistemological point of view, association studies and mechanistic studies reinforce each other—their combined evidential value is more than the sum of the parts. The case study of Auker-Howlett and Wilde (2019), discussed above, shows this reinforcing in action. This epistemological reinforcing can be expected to further reduce the influence of subjectivity.

Practice. So far, we have seen that formalisation offers little scope for addressing malleability, but that we should nevertheless expect EBM+ to be less prone to malleability than EBM because association and mechanistic studies offer a greater quantity and variety of evidence and they reinforce one another. Actual practice supports this claim. Abdin et al. (2019) consider evaluations of amoxicillin as a cause of drug reaction with eosinophilia and systemic symptoms (DRESS). This is a case in which considering mechanistic studies offers to be particularly promising: there is too little evidence from association studies for an informative EBM evaluation, because these adverse drug reactions are extremely rare and can take many years to materialise. The authors apply both the EBM+ approach and the *E*-synthesis approach that we encountered above. They demonstrate inter-tool agreement: they show that the two approaches yield similar conclusions, namely that mechanistic evidence lends further support to the claim that amoxicillin is a cause of DRESS, but not enough support to establish the claim.

We have already seen that IARC's current approach to evaluation is very close to the EBM+ approach. Anecdotally at least, there seems to be little inter-assessor variability in assessments of mechanistic studies carried out by the mechanistic subgroup of an IARC evaluation. There is certainly no evidence that subjectivity has a significant influence on overall evaluations. Simple structural features of IARC's methodology help to avoid malleability. Firstly, potential financial conflicts of interest are taken very seriously. (It sometimes happens that an assessor is removed from a working group in the middle of a review meeting, when a potential conflict of interest is found.) Second, scientists are not permitted to evaluate their own studies, so there is less scope for intellectual conflicts of interest to influence proceedings. Third, the IARC secretariat work very hard to ensure consistency across evaluations, by ensuring that assessors are aware of normal standards by which judgements of strength of evidence are made, and by adopting a very formu-

laic procedure for integrating subgroup assessments in order to converge upon an overall assessment of carcinogenicity. Malleability is kept in check in practice.[8]

Malleability is an understandable worry, given problems faced by EBM in the assessment of association studies. However, we have seen that there are several good reasons for thinking that a move to EBM+ will mitigate, rather than amplify, the effects of subjective choices during the assessment procedure.[9] While more clearly needs to be done to test for malleability in EBM+, concerns about malleability are certainly not grounds for choosing EBM over EBM+.

§4 Discussion

We have seen that EBM+ can be defended against two charges: that is it unfeasible and that it is malleable.

While there is limited evidence of feasibility in the areas of intervention and disease assessment, there is good evidence of feasibility arising from IARC practice in exposure assessment. Moreover, the feasibility of EBM+ in one area of practice supports its feasibility in other areas: assessing causality is a general problem that transcends these rather arbitrary distinctions between kinds of practice. Indeed, the lessons learned here apply beyond medicine. For example, Shan and Williamson (2020) argue that the basic epistemological framework underpinning EBM+, namely a particular form of epistemological pluralism, can also be applied to the social sciences, including to evidence-based policy (EBP), which leads to EBP+, and to basic social sciences research, where the framework can be viewed as providing foundations for mixed methods research.

With regard to malleability, we have seen that there are general epistemological reasons for thinking that subjective influences are likely to be diminished by considering mechanistic studies alongside association studies, and that actual practice suggests that malleability is not a substantial problem for EBM+. (Again, the general epistemological considerations carry over to the social sciences.) This is not to suggest that EBM+ eradicates the need for personal judgement—judgements of quality of study, for example, are required by both EBM and EBM+. The claim is that by considering mechanistic studies in addition to association studies, one has more evidence to go on, more varied evidence, and evidence that makes up for the deficiencies of other evidence, so there is less scope for any malleability with respect to individual judgements to influence the final assessment of causality.

Stegenga (2018) appeals to the malleability of EBM to argue for *medical nihilism*: the claim that almost all medical interventions are ineffective. His argument can be put roughly as follows: medical interventions are approved on the basis of EBM assessments; EBM assessments are riddled with problems, such as malleability,

[8] Far from being malleable, Williamson (2019b) argues that, if anything, IARC evaluations are not flexible enough to cope with exceptional cases.

[9] Recall that Howick (2019, p. 178) criticises EBM+ for not focussing on the problem of financial biases. We have seen above that the diversity of evidence considered by EBM+ helps to ameliorate this problem. Coupling that progress with a healthy scepticism towards research carried out by researchers or organisations with potential financial conflicts of interest, and a strategy for avoiding evidence appraisers with potential financial conflicts of interest, goes a long way towards solving the problem. Thus EBM+ is on a better footing than EBM with respect to financial biases.

which weaken the link between effectiveness and approval; so we should have low confidence that an approved medical intervention is effective.

What should we make of Stegenga's argument? At least two considerations urge caution. First, Stegenga appeals to Bayesianism to formalise his argument. Now, as we have seen, Bayesian explications of scientific confirmation are themselves prone to malleability: one can easily take issue with Stegenga's claims about the probabilities that feature in his explication and reach different conclusions, as Gillies (2019b) explains. Thus there is a sense in which Stegenga's focus on malleability is self-undermining.

Second, one can take issue with the first premise of Stegenga's argument: that medical interventions are approved on the basis of EBM assessments. Certainly, intervention approval panels almost always *claim* to come to their judgements by means of the methods of present-day EBM. Certainly, the methods of present-day EBM—especially the assessment of RCTs and meta-analyses and systematic reviews of RCTs—*inform* their judgements. But their judgements tend not to wholly comply with the principles of present-day EBM. Typically, analyses of RCTs, meta-analyses and systematic reviews are presented to approval panels, and then a general informal discussion ensues—a discussion which often encompasses mechanisms of action, compliance and adverse effects—before a judgement is made. Panel members give opinions about the plausibility of the underlying mechanisms, and this plausibility informs the resulting judgement about whether the intervention should be approved. This part of the process does not accord with EBM, which, as we noted in §1, holds that mechanistic reasoning and expert opinion should be given little or no weight in comparison to RCTs. So the approval process does not altogether follow the precepts of EBM.

That intervention approval departs from the precepts of EBM undermines Stegenga's argument for medical nihilism. This is because the point of departure is with regard to mechanistic hypotheses, and this shifts the approval process in the direction of EBM+. That in practice approval panels explicitly evaluate association studies and consider mechanistic hypotheses in an implicit, common-sensical way lends some confidence to the process, because it is a step in the direction of EBM+. In order to cast doubt on the approval process, Stegenga would need to undermine what takes place in practice, i.e., this common-sensical hybrid, which we might call EBM±.

This is not to put the approval process beyond criticism. It would obviously be better if, instead of an informal discussion of mechanisms guided by the expertise and interests of the panel members who happen to be on the panel, mechanistic evidence were systematically scrutinised and its assessment systematically integrated with that of correlational evidence. As we saw in §1, the guiding principle underlying both EBM and EBM+ is that it is largely by making the evidence and the appraisal process explicit and systematic that one can improve the reliability of resulting judgements; there is clearly much more to be done here with respect to the practice of intervention approval.[10] However, the further one moves towards EBM+, the more confidence one can have in judgements made on the basis of evidence appraisals.

[10]Moving further towards EBM+ requires some changes to the infrastructure of evidence evaluation, to ensure that mechanistic evidence is systematically considered and integrated with evidence of correlation. But, as we have seen in the case of IARC, this infrastructure is of the same kind as that required to evaluate association studies. See Aronson et al. (2018, §12) on this point, in relation to the assessment of mechanistic evidence for drug approval.

Acknowledgements

I am very grateful to Michael Wilde for many helpful comments and to the following organisations for allowing me to observe evidence appraisal meetings: the International Agency for Research on Cancer (IARC); the UK Medicines and Healthcare products Regulatory Agency (MHRA); and the UK National Institute for Health and Care Excellence (NICE). This research was supported by the Leverhulme Trust grant RPG-2019-059 and the UK Arts and Humanities Research Council (AHRC) grant AH/M005917/1.

Bibliography

Abdin, Y., Auker-Howlett, D. J., Landes, J., Mulla, G., Jacob, C., and Osimani, B. (2019). Assessing the mechanistic evidence assessors E-Synthesis and EBM+: A case study of amoxicillin and drug reaction with eosinophilia and systemic symptoms (DRESS). *Current Pharmaceutical Design*, 25(16):1866–1880.

Andreoletti, M. and Teira, D. (2019). Rules versus standards: What are the costs of epistemic norms in drug regulation? *Science, Technology, & Human Values*, 44(6):1093–1115.

Aronson, J. K., La Caze, A., Kelly, M. P., Parkkinen, V.-P., and Williamson, J. (2018). The use of mechanistic evidence in drug approval. *Journal of Evaluation in Clinical Practice*, 24(5):1166–1176.

Auker-Howlett, D. and Wilde, M. (2019). Reinforced reasoning in medicine. *Journal of Evaluation in Clinical Practice*, published online. https://doi.org/10.1111/jep.13269.

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

Conradi, U. and Joffe, A. R. (2017). Publication bias in animal research presented at the 2008 Society of Critical Care Medicine conference. *BMC Research Notes*, 10(262):1–11.

Dorling, J. (1975). Review of *The Structure of Scientific Inference* by M.B. Hesse. *The British Journal for the Philosophy of Science*, 26(1):61–71.

Dorling, J. and Edgington, D. (1976). The applicability of Bayesian convergence-of-opinion theorems to the case of actual scientific inference. *The British Journal for the Philosophy of Science*, 27(2):160–161.

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press, Pittsburgh.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. MIT Press, Cambridge MA.

Frank, C., Faber, M., and Stark, K. (2016). Causal or not: applying the Bradford Hill aspects of evidence to the association between Zika virus and microcephaly. *EMBO Molecular Medicine*, 8(4):305–307.

Fugh-Berman, A. (2013). How basic scientists help the pharmaceutical industry market drugs. *PLoS Biology*, 11(11):1–5.

Gillies, D. (2019a). Holman's criticisms of EBM+. ResearchGate. http://dx.doi.org/10.13140/RG.2.2.15291.77608.

Gillies, D. (2019b). Should we distrust medical interventions? *Metascience*, 28(2):273–276.

Green, S. B. (2015). Can animal data translate to innovations necessary for a new

era of patient-centred and individualised healthcare? Bias in preclinical animal research. *BMC Medical Ethics*, 16(53):1–14.

Guyatt, G., Cairns, J., Churchill, D., Cook, D., Haynes, B., Hirsh, J., Irvine, J., Levine, M., Levine, M., Nishikawa, J., Sackett, D., Brill-Edwards, P., Gerstein, H., Gibson, J., Jaeschke, R., Kerigan, A., Neville, A., Panju, A., Detsky, A., Enkin, M., Frid, P., Gerrity, M., Laupacis, A., Lawrence, V., Menard, J., Moyer, V., Mulrow, C., Links, P., Oxman, A., Sinclair, J., and Tugwell, P. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA*, 268(17):2420–2425.

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300.

Holman, B. (2019). Philosophers on drugs. *Synthese*, 196(11):4363–4390.

Howick, J. (2011). Exposing the vanities—and a qualified defence—of mechanistic evidence in clinical decision-making. *Philosophy of Science*, 78(5):926–940. Proceedings of the Biennial PSA 2010.

Howick, J. (2019). Exploring the asymmetrical relationship between the power of finance bias and evidence. *Perspectives in Biology and Medicine*, 62(1):159–188.

IARC (2019a). *IARC Monographs on the Evaluation of Carcinogenic Hazards to Humans: Preamble*. Lyon. https://monographs.iarc.fr/wp-content/uploads/2019/01/Preamble-2019.pdf.

IARC (2019b). *Styrene, Styrene-7,8-oxide, and Quinolene*, volume 121 of *IARC Monographs on the Evaluation of Cacinogenic Risks to Humans*. International Agency for Research on Cancer. https://publications.iarc.fr/_publications/media/download/6060/5894fec08d186b0eb1a24cfa93db2cd97dc2eb2c.pdf.

Krauer, F., Riesen, M., Reveiz, L., Oladapo, O. T., Martínez-Vega, R., Porgo, T. V., Haefliger, A., Broutet, N. J., Low, N., and Group, W. Z. C. W. (2017). Zika virus infection as a cause of congenital brain abnormalities and Guillain-Barré syndrome: Systematic review. *PLOS Medicine*, 14(1):1–27.

La Caze, A. (2019). Better evaluating mechanisms in medicine. Book review: Evaluating evidence of mechanisms in medicine. *Journal of Evaluation in Clinical Practice*, 25(6):1228–1231.

Landes, J. (2020). Variety of evidence. *Erkenntnis*, 85:183–223.

Landes, J., Osimani, B., and Poellinger, R. (2018). Epistemology of causal inference in pharmacology: Towards a framework for the assessment of harms. *European Journal for Philosophy of Science*, 8(1):3–49.

NICE (2010). Peginterferon alfa and ribavirin for the treatment of chronic hepatitis C. National Institute for Health and Care Excellence. Technology appraisal guidance TA200, https://www.nice.org.uk/guidance/ta200.

OCEBM Levels of Evidence Working Group (2011). The Oxford 2011 levels of evidence. Oxford Centre for Evidence-Based Medicine, https://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf.

Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., Norell, C., Russo, F., Shaw, B., and Williamson, J. (2018). *Evaluating evidence of mechanisms in medicine: principles and procedures*. Springer, Cham, Switzerland.

Posadzki, P. P., Bajpai, R., Kyaw, B. M., Roberts, N. J., Brzezinski, A., Christopoulos, G. I., Divakar, U., Bajpai, S., Soljak, M., Dunleavy, G., Jarbrink, K., Nang, E. E. K., Soh, C. K., and Car, J. (2018). Melatonin and health: an umbrella review of health outcomes and biological mechanisms of action. *BMC Medicine*, 16(1):18.

Rasmussen, S. A., Jamieson, D. J., Honein, M. A., and Petersen, L. R. (2016). Zika virus and birth defects—reviewing the evidence for causality. *New England Jour-*

nal of Medicine, 374(20):1981–1987.

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., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72.

Samet, J. M., Chiu, W. A., Cogliano, V., Jinot, J., Kriebel, D., Lunn, R. M., Beland, F. A., Bero, L., Browne, P., Fritschi, L., Kanno, J., Lachenmeier, D. W., Lan, Q., Lasfargues, G., Curieux, F. L., Peters, S., Shubat, P., Sone, H., White, M. C., Williamson, J., Yakubovskaya, M., Siemiatycki, J., White, P. A., Guyton, K. Z., Schubauer-Berigan, M. K., Hall, A. L., Grosse, Y., Bouvard, V., Benbrahim-Tallaa, L., Ghissassi, F. E., Lauby-Secretan, B., Armstrong, B., Saracci, R., Zavadil, J., Straif, K., and Wild, C. P. (2020). The IARC Monographs: Updated procedures for modern and transparent evidence synthesis in cancer hazard identification. *JNCI: Journal of the National Cancer Institute*, 112(1):1–8.

Shan, Y. and Williamson, J. (2020). Applying Evidential Pluralism to the social sciences. *Under review*.

Solomon, M. (2015). *Making medical knowledge*. Oxford University Press, Oxford.

Stegenga, J. (2018). *Medical nihilism*. Oxford University Press, Oxford.

Swaen, G. and van Amelsvoort, L. (2009). A weight of evidence approach to causal inference. *Journal of Clinical Epidemiology*, 62(3):270–277.

Teira, D. and Reiss, J. (2013). *Causality, Impartiality and Evidence-Based Policy*, pages 207–224. Springer Netherlands, Dordrecht.

Weed, D. L. (2005). Weight of evidence: A review of concept and methods. *Risk Analysis*, 25(6):1545–1557.

Wilde, M. and Parkkinen, V.-P. (2019). Extrapolation and the Russo–Williamson thesis. *Synthese*, 196(8):3251–3262.

Williamson, J. (2018). Establishing the teratogenicity of Zika and evaluating causal criteria. *Synthese*, https://doi.org/10.1007/s11229-018-1866-9.

Williamson, J. (2019a). Establishing causal claims in medicine. *International Studies in the Philosophy of Science*, 32(2):33–61.

Williamson, J. (2019b). Evidential Proximity, Independence, and the evaluation of carcinogenicity. *Journal of Evaluation in Clinical Practice*, 25(6):955–961.

Zhang, J., Liu, H., and Seidenfeld, T. (2018). Agreeing to disagree and dilation. *International Journal of Approximate Reasoning*, 101:150–162.