Studies in History and Philosophy of Science 91 (2022) 288–295

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Evidence of effectiveness

Jacob Stegenga

Department of History and Philosophy of Science, University of Cambridge, United Kingdom

ARTICLE INFO

Keywords:
Evidence
Inference
Randomised trials
Mechanisms
Evidence-based medicine
Evidence-based policy

ABSTRACT

There are two competing views regarding the role of mechanistic knowledge in inferences about the effectiveness of interventions. One view holds that inferences about the effectiveness of interventions should be based only on data from population-level studies (often statistical evidence from randomised trials). The other view holds that such inferences must be based in part on mechanistic evidence. The competing views are local principles of inference, the plausibility of which can be assessed by a more general normative principle of inference. Bayesianism tells us to base inferences on both the 'likelihood' and the 'prior'. The likelihood represents statistical evidence. One influence on the prior probability of a hypothesis like '*d* causes *x*' is mechanistic knowledge of how *d* causes *x*. Thus, reasoning about such inferences by appealing to both statistical and mechanistic evidence is vindicated by our best general theory of inference. The primary contribution of this paper is to assess the merits and weaknesses of the arguments on both sides of the debate, using the Bayesian framework. This analysis lends support to those who argue that we should base our causal inferences about interventions in part on mechanistic evidence.

1. Introduction

There are two opposed views regarding what kinds of evidence are needed to warrant causal inferences in the medical and social sciences. This paper appeals to a general prescriptive theory of scientific inference—Bayesianism—to explain why the arguments that have been proposed for both sides can seem compelling, and under what conditions each of the arguments is in fact persuasive. This ultimately vindicates the position that holds that causal inferences in these domains should be based in part on mechanistic reasoning.

This debate is about the kinds of evidence required for justifying inferences about the effectiveness of interventions, such as the effects of a new pharmaceutical or a poverty-reduction program. One radical view is the *statisticalista* thesis, which holds that causal inferences in medicine and the social sciences should be based only on evidence from randomised controlled trials or other population-level studies (call this kind of evidence 'statistical evidence'). If (and only if) evidence from a population-level study such as a randomised trial suggests that intervention *d* has effect *x*, then statisticalistas hold that we should believe, or at least act as if, *d* causes *x*. Statisticalistas maintain that mechanistic evidence is typically unreliable and ought not be considered in these contexts, but the key point of this position is that mechanistic evidence simply need not be considered.

The opposing view is the *mechanista* thesis, which holds that such causal inferences must be based, at least in part, on mechanistic evidence. Mechanistic evidence for *d* causes *x* is evidence of causal mechanisms linking *d* to *x*, or evidence of mechanisms that could otherwise modulate the *d*-*x* causal relation (Illari, 2011; Claveau, 2012). Mechanistas hold that statistical evidence is insufficient to warrant causal inferences, and that mechanistic evidence is necessary. If statistical evidence suggests that *d* causes *x*, and we have no mechanistic evidence pertaining to this alleged causal relation, then a strong mechanista position would hold that we are not justified in believing that *d* causes *x*. A radical mechanista position would hold that mechanistic evidence alone can be sufficient to warrant causal inferences about interventions, yet most mechanistas hold that both statistical evidence and mechanistic evidence on their own are insufficient for warranting causal inferences, but both are necessary and when taken together are sufficient.[1] In short, the dispute between statisticalistas and mechanistas is about whether or not mechanistic evidence is necessary for causal inference—statisticalistas say it is not, while mechanistas say it is (statisticalistas further claim that mechanistic evidence is typically unreliable).

E-mail address: jms303@cam.ac.uk.

[1] The kinds of causal hypotheses of interest in the debate tend to be type-level claims about a population (such as 'this drug lowers blood pressure') (Williamson, 2019), though some disputants argue that the evidential requirements for token-level claims about individuals plausibly differ from the evidential requirements for type-level claims.

https://doi.org/10.1016/j.shpsa.2022.01.001
Received 31 March 2021; Received in revised form 2 January 2022

0039-3681/© 2022 Elsevier Ltd. All rights reserved.

These two opposing views are adopted by various philosophers, scientists, and regulators. For instance, the so-called Russo-Williamson thesis is named after two philosophers who have defended the mechanista position (Russo & Williamson, 2007). Conversely, Howick has criticised the mechanista position and has defended the statisticalista thesis (though his position is more nuanced, and is in fact not far from the position defended here (Howick, 2011b)).[2] The evidence-based medicine and evidence-based policy movements have explicitly codified the statisticalista view into their methodological principles. The European Medicines Agency bases its regulatory approval of new drug applications on the statisticalista view, because the standard simply requires evidence from randomised trials. Conversely, the International Agency for Research on Cancer bases its assessment of chemical carcinogenicity on the mechanista view (Leuridan, 2011) (although the causal hypotheses studied by IARC are about toxicity of chemicals rather than effectiveness of interventions, the evidential issues are similar). So the question addressed in this paper—what kind of evidence is required for causal inference in domains like medicine and social science—is an unsettled one for philosophers, scientists, and regulators.[3] Moreover, the stakes are high, since the causal hypotheses in these domains are often of great practical importance.

Both statisticalistas and mechanistas defend their positions with relatively persuasive arguments that involve appealing to methodological reliability and historical cases. This generates an awkward tension, since the two views cannot both be generally correct. The two positions—the statisticalista thesis and the mechanista thesis—are each constituted by local principles of inference: both views state that to make such-and-such kinds of inference in such-and-such contexts, one should rely on such-and-such kinds of evidence. The ambition of this paper is to assess these two competing local principles of causal inference by appealing to a more general, prescriptive, and independently warranted theory of scientific inference, namely, Bayesianism. This is, to the best of my knowledge, the first time that this debate has been cast in these terms. With this general principle of inference in hand I articulate the conditions under which the key arguments for both sides of the debate are compelling. The contribution of this article is not merely to offer a formal representation of the key arguments in the debate, but rather, the contribution is to use the formal representation to shed light on the merits and weaknesses of the main arguments for both sides in the debate, and to articulate the conditions under which the arguments apply. While others have offered critical assessments of some of the relevant arguments (e.g. Marchionni & Reijula, 2019; Williamson, 2019), the approach taken here provides a novel re-framing of the debate. Perhaps most saliently, the Bayesian emphasis on considering the prior probability of a hypothesis when making an inference serves to highlight the importance of background knowledge, which plausibly includes considerations of relevant mechanisms underlying the causal relation in question.

I identify two key arguments for both the statisticalista view (§2) and the mechanista view (§3), and with the help of the formal representation of the arguments, I show under what conditions each of the arguments is, and is not, persuasive (§4). The formal representation here serves as an aid to uncover background premises in each argument. Ultimately, this provides some vindication of a modest mechanista position.

2. Statisticalista

The statisticalista holds that evidence from population-level studies (statistical evidence) is necessary and sufficient for making causal inferences about interventions in practical domains such as medicine and social policy. The view holds that such evidence should typically come from randomised trials or meta-analyses of such trials. The statisticalista view has been central to the tenets of evidence-based medicine, evidence-based policy, and regulatory policies.

The evidence-based medicine and evidence-based policy movements are built around the idea that statistical evidence is necessary for warranting causal hypotheses about interventions, and that evidence of other types, including mechanistic evidence, is unreliable. The epistemic status of randomised trials versus other kinds of population-level studies has been the subject of very active debate, specifically regarding the value of random allocation of subjects to experimental groups.[4] The evidence-based movements hold that randomised population-level studies are superior to non-randomised population-level studies, and, for inferences about the effectiveness of interventions, any population-level study (randomised or not) is superior to inferences about the effectiveness of the intervention based on mechanistic evidence. This view is sometimes represented by so-called 'evidence hierarchies', which rank-order sources of evidence according to their alleged reliability: randomised trials are placed at the top, non-randomised population-level studies are placed lower, and mechanistic evidence is somewhere near the bottom.[5] It has long been a motif of evidence-based medicine and policy that mechanistic evidence is generally unreliable for informing inferences about the effectiveness of interventions.

This view has been influential on policy. For example, the standard for new drug approval is that two randomised trials which suggest that a new drug is better than placebo are necessary and sufficient for approval of this drug.

In what follows I describe two key arguments for the statisticalista view.

2.1. Evidence-based argument

Consider: if statistical evidence from the most reliable population-level studies suggests that d causes x, then what more do we need in order to infer that d causes x? The most responsible thing for us to believe is that d causes x; that is what the evidence says, and we should believe the evidence (Howick, 2011a). I will call this the evidence-based argument for the statisticalista.

For example, in 1753 when James Lind found that eating citrus could help treat scurvy, scientists had no understanding of how citrus could do this. Lind performed a trial—this has been called the first or at least among the first of medical trials—in which some sailors with scurvy received citrus juice and other sailors received other interventions (vinegar, cider, sea water …). The sailors who received citrus juice recovered from their scurvy, while the other sailors did not. This was population-level statistical evidence. Since there was no knowledge of vitamins, let alone vitamin deficiency, there could not possibly have been mechanistic understanding of the pathophysiology of scurvy or of citrus juice as an intervention.

To consider a more recent example, some of the most careful meta-analyses of randomised trials of antidepressants suggest that such drugs have very tiny effect sizes: a reduction of less than two points on a fifty-point depression severity scale, compared with placebo. The evidence-based argument encourages us to ask what our best estimate should be

[2] Compare: "mechanistic reasoning is *not* necessary to establish causal claims" (Howick, 2011a, emphasis in original) with "To establish causal claims, scientists need the mutual support of mechanisms and dependencies" (Russo & Williamson, 2007). See Grüne-Yanoff (2016) for support of the latter position in behavioural policy.

[3] On the notion of effectiveness, see Stegenga (2015a).

[4] For instance, a foundational article for the evidence-based medicine movement states that one should "discard at once all articles on therapy that are not about randomised trials" (Department of Clinical Epidemiology and Biostatistics, 1981). A textbook of evidence-based medicine states that "if a study wasn't randomised, we suggest that you stop reading it and go on to the next article in your search" (Straus et al., 2005).

[5] See, for example, the website of the Oxford Centre for Evidence-Based Medicine, which puts randomised trials near the top of their evidence hierarchy, while inferences "based on physiology" are placed at the bottom.

of the effectiveness of these drugs. An average reduction of one point on the scale? Ten points? If your view is that once we have gathered all of the best population-level evidence about antidepressants in a careful meta-analysis, our estimate of the effectiveness of antidepressants should be close to whatever the result of the meta-analysis is, then your intuition supports the evidence-based argument. If that is not your view, then your view is at odds with the best evidence, goes this argument, and thus at the very least you must offer a reason as to why your estimate of the effectiveness of antidepressants should not be close to whatever the result of the meta-analysis is, and at worst you are not being a good empiricist.

2.2. Trouble-in-the-box argument

Another argument in favour of the statisticalista view is that our knowledge of mechanisms is often incomplete and can lead us astray. This is especially so when mechanistic reasoning is based on faulty physiological theories, or the target system about which we are reasoning is complex. Since mechanistic reasoning is so unreliable, this argument goes, we should base our causal inferences only on statistical evidence. To take an extreme example, under humoral theory psychosis was thought to be the result of excess blood and yellow bile, and thus, based on mechanistic reasoning, the standard treatment for psychosis was bloodletting. Since we now know that bloodletting was at best useless and very likely harmful for such people, the mechanistic reasoning in this case was unreliable. In general, claims the statisticalista, the mechanistic details about how interventions modulate their targets should be thought of as black boxes, because their causal complexity so easily misleads us when we try to understand them. This is the trouble-in-the-box argument for the statisticalista. It is a staple in evidence-based medicine and philosophical commentators—(Howick, 2011a), for example, offers several persuasive examples to illustrate the argument, and (Broadbent, 2011) endorses it.

An example that many contributors to this debate have appealed to is the CAST trial, performed in the 1980s to test if a class of drugs, antiarrhythmics, could lower mortality following a heart attack. The reasoning was that some people die in the months following a heart attack, and many of these deaths are the result of cardiac arrhythmias, so if we could suppress arrhythmias, we could reduce the number of post-heart attack deaths. The results from the trial showed the opposite: people on antiarrhythmics in fact had a higher rate of mortality following heart attacks. Statisticalistas take this example to show that the mechanistic reasoning which motivated the trial was unreliable, and this unreliability was demonstrated by the statistical results of the trial (Howick 2011). Of course, one will only be persuaded by an example like this if one thinks that the statistical results from a trial are decisive for the particular causal inference, but that is precisely what is in question in this debate.

A more fundamental reason that some statisticalistas cite to explain the unreliability of mechanistic reasoning is the complexity of target systems. In medicine target systems are typically complex because of evolutionary history. In social policy target systems are complex because human behaviour and social relations are complicated. Andersen, for instance, notes "the complex, evolved, layered causal structure of the human body, which results in mechanisms that frequently are non-modular and involve 'hidden' causal relationships" (Andersen, 2012).[6] Modularity of causal systems exists when one part of the system can be intervened upon without modifying other parts of the system. In complex causal systems (which are typical in the medical and social sciences), causal systems are not modular.

Such complexity generates problems for mechanistic reasoning for both epistemic and ontological reasons. Epistemically, the problem is that in many cases we do not have complete descriptions of the pertinent mechanisms, and so reasoning mechanistically is unreliable. If an unknown part of a mechanism includes a set of causal relations which form a 'negative feedback loop', then the actual outcome of an intervention could be the exact opposite of that which is predicted by mechanistic reasoning. Ontologically, the problem is that in many cases interventions in one part of a mechanism can lead to reconfigurations of the causal structure of the mechanism elsewhere, which can modulate the effects of the intervention.[7]

3. Mechanista

The mechanista holds that statistical evidence is insufficient for causal inferences in these domains. Mechanistas argue that such inferences should be based, at least in part, on mechanistic evidence. One mechanista view holds that mechanistic evidence and statistical evidence must be taken together to warrant causal inferences (thereby granting that statistical evidence is at least necessary for such inferences). Perhaps the most prominent statement of the mechanista position is that of Russo and Williamson (2007): "To establish causal claims, scientists need the mutual support of mechanisms and dependencies" (where 'dependencies' refers to statistical evidence from population-level studies). A stronger mechanista position holds that mechanistic evidence is at least sometimes sufficient on its own to warrant causal hypotheses; Claveau (2012), for example, argues that in the social sciences causal claims can be established by mechanistic evidence alone.

Precisely what mechanistic evidence means, and what the mechanista position requires, has been a matter of some discussion.[8] I adopt a relatively standard view that mechanistic evidence is evidence of causal mechanisms linking, or otherwise modulating the relation between, d and x. Some contributions to the mechanista position have argued that what is important is evidence *of* the existence of a mechanism linking d to x, rather than evidence from a particular kind of method often used to generate mechanistic evidence. On this view, statistical evidence from a population-level trial can itself provide evidence of a mechanism.[9] Of course, if there is any statistical evidence that d causes x, that alone is some evidence of the existence of a mechanism linking d to x, however minimal. If the mechanista view were merely that one must have evidence of mechanisms, and such evidence can be statistical evidence from randomised trials, then the position would be methodologically similar to the statisticalista view. Yet, as already noted, the mechanista view is routinely articulated as offering a methodological competitor to the statisticalista view (see, e.g., Williamson, 2019). Marchionni and Reijula (2019) offer a compelling articulation of the mechanista view: they argue

[6] Anderson argues that despite the problems of mechanistic reasoning when applied to inferences about the effectiveness of interventions, mechanistic reasoning nonetheless has a role to play in making recommendations to individual patients based on population-level evidence. This view has had some impact; see for example the discussion in Marchionni and Reijula (2019).

[7] A related, practical issue is that medical science has well-developed guidelines for assessing the quality of statistical evidence, but assessing the quality of mechanistic evidence has not been as well-developed. On the other hand, Stegenga (2015b) argues that different guidelines for assessing the quality of statistical evidence can wildly disagree, in part because the various guidelines are not grounded on established principles of scientific inference. Conversely, Parkkinen et al. (2018) is a contribution toward developing standards for assessing mechanistic evidence.

[8] Mechanisms themselves have been the subject of a voluminous literature; see, for example, (Machamer et al., 2000) and (Bechtel & Abrahamsen, 2005).

[9] For example, Clarke et al. (2014) claim: "what is required is evidence of two different sorts of things—correlation and mechanisms—not two different kinds of evidence (Illari, 2011). Indeed, a single item of evidence can be evidence of both correlation and mechanisms. For instance, in principle a well devised and well conducted RCT can on its own provide evidence for a causal claim, since it can provide evidence of correlation, and, if in the circumstances other explanations of this correlation are sufficiently implausible, it can also provide evidence that there is some underlying mechanism linking the putative cause and the putative effect that can account for the correlation." (343).

that mechanistic evidence about a hypothesis *d* causes *x* involves information about the causal pathway from *d* to *x,* and the relevant causes constituting the mechanism can be at different levels or physical scales, and the information can be gleaned from a diverse range of methods (see also Illari, 2011).

Sometimes mechanistas argue that to extrapolate statistical evidence reliably we need to appeal to mechanistic evidence. Since the causal inferences in question are about interventions to be used in practical contexts, we need to be able to extrapolate from the confined contexts in which population-level statistical evidence is typically produced (experimental contexts) to the target contexts in which such interventions will be deployed. There are many ways such extrapolations can fail, and thus to extrapolate reliably we must think through how interventions might operate, and fail to operate, in target contexts. This involves reasoning mechanistically about the intervention and the target (for versions of this argument, see (Steel, 2007), (Cartwright, 2012), and (Fuller, 2015)). My focus here, however, is on causal inference per se rather than extrapolation.

Two main arguments for the mechanista view are as follows.

3.1. Know-how argument

Consider: if we know how *d* causes *x*, then it follows that we know that *d* causes *x*, and we do not further need any other kind of evidence, such as statistical evidence, to establish that *d* causes *x*. I will call this the know-how argument for the mechanista view. A silly but striking example is sometimes used to illustrate this argument: we know how parachutes slow one's fall, and so we know that they do this, and it would be absurd, when testing the effectiveness of parachutes to slow one's fall, to demand statistical evidence for this inference—a population-level study comparing parachute-wearers to non-parachute-wearers testing the capacity of parachutes to slow one's fall (Smith & Pell, 2003). (Of course, now we have some statistical evidence about the effectiveness of parachutes, which comes from a series of individuals using parachutes, but understood as population-level statistical evidence, this series of cases is unblinded, uncontrolled, and non-randomised, and so hardly fits the mould of quality statistical evidence.)

Sometimes we have very good reason to believe that *d* does not cause *x*. In such cases the know-how argument is just as persuasive, though in contrast with cases like the parachute case, the conclusion for this other class of cases is obviously that *d* does not cause *x*. There are some comic examples of this in recent medical research. For instance, a researcher performed a randomised trial on the efficacy of retroactive intercessory prayer (Leibovici, 2001). The subjects in this trial were patients who had been in a hospital several years prior to the trial. The researchers randomly allocated some of the patients to the intervention group and some to the control group. The patients in the intervention group received intercessory prayer, while the patients in the control group received nothing. Various outcomes were measured and the researchers found that patients who received retroactive intercessory prayer had slightly better outcomes than patients who received no intervention. Of course, the implication of this trial is precisely the opposite of believing in the effectiveness of the intervention, and the mechanista has an explanation for why that is the case (the know-how argument) while the statisticalista does not.[10]

There are many more serious examples of causal inferences illustrating the know-how argument. For example, we can inform some inferences about the effects of a drug by appealing to knowledge of how that drug modulates our physiology. Consider the class of drugs known as PPAR modulators. These drugs regulate the expression of many genes that influence one's physiological functioning in many parts of the body. So if one's hypothesis is suitably coarse-grained, such as 'this PPAR modulator will have many side effects', then one can, by reasoning mechanistically, be relatively confident in one's hypothesis, since we know that such drugs modulate the expression of genes that influence many aspects of physiology.

3.2. Trouble-in-the-data argument

Another argument in favour of the mechanista view is that statistical evidence is often misleading and can lead us astray. Population-level studies, including randomised trials and meta-analyses, suffer from many biases, so when using statistical evidence to warrant causal inferences such evidence should be supplemented with mechanistic evidence. Mechanistic evidence might also be unreliable (as noted above), but it is unreliable in different ways than statistical evidence is. Thus, at the very least statistical evidence should be supplemented with mechanistic evidence. I will call this the trouble-in-the-data argument for the mechanista thesis.

The retroactive intercessory prayer example, noted above as an illustration of the know-how argument, also serves as an illustration of the trouble-in-the-data argument. A randomised trial published in a reputable medical journal suggested that retroactive intercessory prayer is effective—thus, we must recognise the fallibility of published randomised trials.

Recent work in philosophy of medicine has offered many more realistic and detailed elucidations and illustrations of this argument. Population-level statistical evidence comes from randomised trials and meta-analyses that are shot through with biases, rendering such evidence very often unreliable—for compelling illustrations of this, see (Biddle, 2007), Stegenga (2011), (Jukola, 2017), (Pinto, 2015), and (Goldacre, 2012).

Notice that the know-how argument is the converse of the trouble-in-the-box argument, and that the trouble-in-the-data argument is the converse of the evidence-based argument. This suggests that both sides are too strong, which motivates a middle position.

4. Resolution

Both the statisticalista thesis and the mechanista thesis are local principles of causal inference, the plausibility of which can be assessed by appealing to a general prescriptive principle of inference. The leading general account of scientific inference is Bayesianism ((Howson & Urbach, 1989), (Sprenger, 2014); for some nuance, see (Sober, 2008)). Suppose that H is a hypothesis about a causal relation like *d* causes *x*. To keep the model simple, E will represent only new population-level statistical evidence relevant to H (we will see in a moment how this model incorporates mechanistic evidence). Bayesians hold that our confidence in H after we receive some new E can be represented as the conditional probability P(H|E), which is, by Bayes' Theorem, equivalent to:

$$P(E|H)P(H) / P(E)$$

So, to infer P(H|E)—the posterior, which is a measure of the final degree of confirmation of our hypothesis after receiving the evidence—not only do we need to take into account P(E|H)—the likelihood, which represents the extent to which the hypothesis fits with or explains the statistical evidence—but we need to take into account P(H)—the prior probability of the hypothesis—and we also need to take into account the expectancy of the evidence, P(E). One influence on the prior probability of a hypothesis like *d* causes *x* is knowledge of how *d* causes *x* (another influence on the prior is past statistical evidence). For example, if *d* is the administration of a drug, an influence on one's prior probability that *d* causes *x* should be evidence pertaining to how *d* is thought to act upon the pathophysiological basis of the disease being intervened upon to bring about *x*. To be clear, nothing about Bayes' Theorem itself suggests that the two kinds of evidence should be represented this way—this

[10] This implication of the retroactive intercessory prayer case might strike one as just as question-begging as the antiarrhythmics example for the statisticalista view, though in the formal representation of the argument below I show how to makes sense of such a case.

is just a convenient representation to articulate the four arguments in question.[11] The formal device allows me to make transparent some premises in each of the arguments which have remained somewhat opaque.

I will use some additional terms in what follows. An 'extremal' probability is either very high or very low. So, an extremal prior implies that P(H) is perhaps less than 0.1 or greater than 0.9. A 'middling' probability is not extremal, and rather is closer to 0.5. So, a middling prior implies that P(H) is, say, somewhere around 0.3 to 0.7. Precisely how high or low a probability must be to be extremal and what precise range of probabilities is implied by a middling prior cannot be decided here, and is anyway immaterial to the arguments that follow.

This appeal to a general theory of inference can explain the persuasiveness of the arguments for both the statisticalista and mechanista positions. The aim of this section is to articulate the conditions under which the four arguments described above (two for each side) are compelling. The upshot is some modest support for the mechanista. As noted, this approach involves a shift of focus, from the principles of causal inference undergirding the two competing views to scientific inference more generally—since causal inference is a species of scientific inference, we can fruitfully assess these two competing views about causal inference by appealing to a more general principle of scientific inference.

4.1. Evidence-based argument, Bayesian Redux

Consider the relative unimportance of mechanistic evidence when we have strong statistical evidence. In the scurvy case the statistical evidence that *d* causes *x* was compelling enough such that no knowledge of mechanisms was necessary to justify an inference that *d* causes *x*: in this case P(E|H) is very high, and in turn the posterior, P(H|E) was high. In other cases a very low P(E|H) entails that the posterior will be very low. In general, in some circumstances an extremal likelihood—when P(E|H) is very high or very low—entails an extremal posterior.

However, an extremal likelihood is not enough to guarantee an extremal posterior. As we saw illustrated by the parachute case and the retroactive intercessory prayer case (analysed further below), an extremal prior can entail that the likelihood, extremal or not, has little influence on the posterior. If the prior probability that H is true is very high, as it is in the parachute case, or very low, as it is in the retroactive intercessory prayer case, then likelihoods have less influence on the posterior than they otherwise would have. So, a condition of application for the evidence-based argument is that the prior must not be extremal.

So, when does the evidence-based argument apply? Statistical evidence has a large impact on the posterior when the likelihood is extremal and the prior is not extremal. In such situations the evidence-based argument applies and the statisticalista view is vindicated.

Howick (2011) takes the evidence-based argument to be a challenge to the mechanista position. In response, Williamson (2019) argues that the Russo-Williamson thesis (an articulation of a mechanista position) can handle such cases, because when there is high quality statistical evidence for a causal relation, that is *also* evidence for the existence of a mechanism underlying that causal relation. Yet, as noted earlier, the mechanista principle that I consider to be most faithful to the arguments motivating the position is that of Marchionni and Reijula (2019), which holds that mechanistic evidence involves information about the underlying causal chain connecting the causal relata. This, in turn, renders Williamson's response to Howick less compelling, since evidence for the mere presence of a statistical correlation, no matter how good that evidence, does not provide specific information about the underlying causal chain connecting the relata. There is a better way to respond to the force of this argument.

As we saw, a case illustrating the evidence-based argument was Lind's experiment, which generated evidence that citrus fruit could alleviate the symptoms of scurvy. Indeed, this is a textbook example for evidence-based medicine. However, perhaps it is not a great example, after all. It is implausible to suppose that the prior for this hypothesis should have been middling—prior to Lind's experiment, that is—because even basic knowledge of vitamins was lacking, let alone vitamin deficiency. This consideration suggests a more general difficulty for the argument.

The evidence-based argument is supposed to be an argument for the statisticalista. Notice a nuance that the scurvy example brings to the fore, and arguably turns the table. As we just saw, a premise of the evidence-based argument is that the prior must not be extremal. One way to know that this is the case for a particular causal hypothesis is to appeal to theoretical or mechanistic knowledge. That is precisely what is involved in assessing whether or not one of the conditions of application of the evidence-based argument is satisfied, namely, whether or not the prior is extremal. So, to know if this statisticalista argument is compelling for any particular case, one must already be engaged in the kind of reasoning that the mechanista urges. What was supposed to be an argument for the statisticalista turns out to be a consideration that offers at least some support for the mechanista.

4.2. Trouble-in-the-box argument, Bayesian Redux

This argument is based on the fact that very often our knowledge of the relevant mechanisms is insufficient for mechanistic reasoning to reliably inform causal inferences. Indeed, the argument asserts that mechanistic reasoning is typically unreliable. This amounts to holding that in typical cases one's prior should not be extremal, at least with respect to the influence of mechanistic reasoning on the prior (that is, putting aside the influence of previous statistical evidence on the prior). That is because, if mechanistic reasoning is typically unreliable, then, for some particular causal hypothesis, based on mechanistic reasoning alone we should have little reason to be very confident in that hypothesis, and little reason to be very unconfident in that hypothesis. The trouble-in-the-box argument holds that mechanistic reasoning does not, in typical cases, provide reasons to be very confident or very unconfident in such a hypothesis. Thus our prior for such a hypothesis ought not be extremal.[12]

This argument is related to the evidence-based argument, insofar as the evidence-based argument only applies when the prior is not extremal. If the trouble-in-the-box argument is sound, then one of the premises of the evidence-based argument holds.

So when is the trouble-in-the-box argument sound? Precisely in those circumstances in which it is true that mechanistic reasoning ought not nudge our priors toward extremal values. That, of course, is a mere formal requirement. I doubt there is anything substantive and general one can say about what those circumstances are. But some basic rules-of-thumb might apply. Howick suggests that we can reliably appeal to mechanistic evidence when our knowledge of the pertinent mechanism is "not incomplete" (2011). So, on this standard, if the mechanistic knowledge were incomplete, then the trouble-in-the-box argument would apply. Taken literally, this 'not incomplete' standard is probably impossible to satisfy in most cases in the medical and social sciences—typically our mechanistic knowledge is incomplete, and so, if this standard were apt, the trouble-in-the-box argument would typically apply, and we could rarely appeal to mechanistic reasoning. This standard is relatively coarse, and with hope others will

[11] Despite this simple representation, it is worth noting that, of course, new mechanistic evidence is routinely being generated for causal hypotheses in medicine.

[12] Strictly speaking, this only follows if there are not other reasons that entail that the prior should be extremal, such as previously-collected and extremely compelling statistical evidence.

develop more refined ways of assessing the reliability of mechanistic evidence.[13]

4.3. Know-how argument, Bayesian Redux

Consider the relative unimportance of statistical evidence when we have compelling mechanistic evidence. In the parachute case, the prior probability that *d* causes *x* is so high that no statistical evidence is needed to justify an inference that *d* causes *x*. In this case P(H) is extremely high, and so P(H|E) can be inferred to be high as well, at least when the statistical evidence is itself not extremal. In the retroactive intercessory prayer case, P(H) is extremely low, and so P(H|E) can be inferred to be low (in this case the statistical evidence was not extremal). In general, if one's prior is extremal, then the posterior is heavily constrained by the prior, and thus the posterior can be little influenced by the likelihood. When we know how *d* causes *x*, we know that *d* causes *x*, regardless of middling statistical evidence pertaining to whether or not *d* causes *x*.

So when does the know-how argument apply? This mechanista argument is compelling insofar as knowledge of the mechanisms by which *d* would cause *x* justifiably entails that the prior probability for the hypothesis that *d* causes *x* is extremal: either very high or very low. When background knowledge justifies extremal priors, the know-how argument applies, and in such situations the mechanista view is compelling. For instance, if a hypothesis is inconsistent with well-substantiated physical theories, then that hypothesis should be assigned an extremely low prior (as in the retroactive intercessory prayer case). Conversely, if a hypothesis is practically entailed by well-substantiated physical theories, then that hypothesis should be assigned an extremely high prior (as in the parachute case).[14] Statistical evidence should still be taken into account if it happens to be available, of course, but its impact on the final degree of confirmation will be anchored by the extremal prior when the know-how argument applies.

The parachute case and the retroactive intercessory prayer case are on opposite ends of a spectrum of plausibility. They are extreme hypotheses with respect to background theory and hence they warrant extremal priors. Most hypotheses in the medical and social sciences are not so extremely plausible or implausible. Indeed, one motivation to test a hypothesis empirically in these practical domains is that the prior probability of that hypothesis is not typically extremal. Conversely, if a hypothesis is virtually guaranteed to be true (like, say, a hypothesis that claimed that the average age of students in an undergraduate classroom is less than sixty years) then scientists would likely not deem it worthy of study; similarly, if a hypothesis is virtually guaranteed to be false (like, say, a hypothesis that claimed that the average age of students in an undergraduate class is less than six years) then scientists also would likely not deem it worthy of study. This suggests that at least one condition on a hypothesis being worthy of pursuit is that it does not have an extremal prior. This is an insufficient condition for pursuit worthiness, of course, because a hypothesis with a middling prior might be uninteresting (like, say, a hypothesis that claimed that the average age of students in an undergraduate class is twenty)—there are, of course, other conditions of pursuit-worthiness, such as intellectual interest or practical importance—but it is, plausibly, a necessary condition on pursuit worthiness.

This in turn suggests that most hypotheses being tested by scientists do not have extremal priors. If so, then the know-how argument would have little applicability. However, the know-how argument doesn't require priors that are virtually zero or one. How extremal does a prior need to be in order for the know-how argument to apply? There is no strict rule here; the influence of a low or high prior will depend on details of particular cases, especially how striking the evidence is.

4.4. Trouble-in-the-data argument, Bayesian Redux

This argument appeals to the unreliability of much research in the medical and social sciences. An empirical method's reliability can be modelled in a Bayesian framework via the expectancy of the evidence, P(E): the more a method is reliable, the lower is the probability of the evidence generated by that method, P(E). Vice versa, the less a method is reliable, the higher is P(E). By Bayes' Theorem, one can easily see that the degree of confirmation of a hypothesis is inversely proportional to P(E). Thus, when the statistical evidence is generated by a less reliable method, the hypothesis receives less confirmation than it would if the evidence was generated by a more reliable method (which is exactly as it should be, of course).

This way of representing the impact of biased methods on confirmation is presented in Stegenga (2018). The argument is as follows. Suppose our E suggests that *d* causes *x*, and one possible explanation of this evidence is the hypothesis that, in fact, *d* causes *x* (call this H_X). E provides some confirmation to H_X. However, if the method that generated E had *i* distinct biases, then E could have resulted from one of those biases, and alternative hypotheses for E are H_{Bi}, one for each bias. Suppose for the sake of simplicity that these are all the biases and that they are mutually exclusive. Then applying the principle of total probability gives us:

$$P(E) = P(E|H_X)P(H_X) + P(E|H_{B1})P(H_{B1}) + P(E|H_{B2})P(H_{B2}).... + P(E|H_{Bi})P(H_{Bi})$$

This allows us to see clearly the impact of bias on the expectancy of the evidence. The terms $P(E|H_n)$ represent how well any particular hypothesis explains E. As the number of biases increases, P(E) increases, because each bias adds a positive term to the right side of the equation. Also, as the prior probability of any bias hypothesis increases, P(E) increases; thus, the more likely it is that a method suffers from any particular bias, P(E) increases. And the more that any particular H_{Bi} explains E, P(E) increases. So, in short, P(E) represents the impact of a method's biases on confirmation—more bias entails higher P(E), and thus lower confirmation (which is, of course, not merely intuitive, but is a practically necessary way of understanding scientific inference).

The impact of evidence on the final degree of confirmation of a hypothesis depends on the ratio of the likelihood, P(E|H), to the expectancy, P(E). The more that P(E|H) exceeds P(E), the more confirmation H receives. (This is true for any plausible measure of confirmation.) So even in cases in which P(E) is high because the method that generated E is biased, E could still provide strong confirmation of H, as long as the likelihood is high. And that would be the case if the evidence were especially supportive of the hypothesis.

So when does the trouble-in-the-data argument apply? Precisely in those circumstances in which the statistical evidence is generated by a method with low reliability (and therefore P(E) is high), and the likelihood is middling. In such circumstances, the statistical evidence offers little confirmation to the hypothesis. Arguably a mistake that the statisticalista view has promulgated is the position that statistical evidence from randomised trials is normally reliable. Moreover, as argued above, the statisticalista position neglects mechanistic evidence. The Bayesian representation provides an independent argument to support the position of those such as Illari (2011) and Russo and Williamson (2007) who maintain that a causal hypothesis can receive some warrant even when the statistical evidence is infused with bias, as long as mechanistic evidence provides some independent warrant for the causal hypothesis. So, even if the trouble-in-the-data argument applies, a causal hypothesis can anyway receive some warrant from mechanistic evidence (this is obvious

[13] See Parkkinen et al. (2018) for a more recent and practical discussion of assessing the quality of mechanistic evidence.

[14] Williamson (2019) also uses these cases to illustrate the importance of mechanistic evidence, and he frames his conclusion in terms of one kind of evidence (statistical) being trumped by another (mechanistic). The approach I take here is consistent with Williamson's, yet while I have described the formal condition (extremal priors) under which such trumping can occur, the Bayesian approach more generally holds that the impact of various forms of evidence on our inference is graded.

given the influence of the prior on the posterior). And, alternatively, if the trouble-in-the-data argument applies and there is little or no mechanistic evidence for the causal hypothesis, then that hypothesis has very thin evidential support.

5. Mind the gap

As we saw above, both sides in this debate appeal to historical examples to illustrate their arguments. Strangely, both sides sometimes appeal to the *same* historical examples in an attempt to prove their point. For instance, both sides appeal to Semmelweis's discovery that doctors washing their hands can lower the rates of childbed fever. Mechanistas argue as follows. The scientific community did not accept Semmelweis's population-level statistical evidence until there was a compelling mechanistic or theoretical explanation of that evidence. At the moment when the scientific community had both statistical evidence and mechanistic evidence, the scientific community accepted Semmelweis's finding. This shows, goes this reasoning, that a widely held standard for scientists is the mechanista principle that to make such a causal inference, not only do we require statistical evidence, we also require mechanistic evidence. Statisticalistas, on the other hand, argue as follows. Had the scientific community accepted Semmelweis's discovery sooner, we would have settled on the truth sooner and many thousands of lives would have been saved. His finding should have been accepted only on the grounds of his population-level statistical evidence, and had this been the case, error would have been avoided and tragedy would have been averted.[15]

These conflicting accounts of the same historical episode illustrate what might be called the 'underdetermination of philosophical theory by historical example.' Both sides appeal to the same historical case and yet they reach opposite philosophical positions. The mechanista argument in cases like this involves appealing to a standard of descriptive accuracy: their local principle of inference is consistent with historical facts in a wide array of cases. The statisticalista argument in cases like this involves appealing to a standard of prescriptive adequacy: their local principle of inference should have been employed in such historical cases (they claim), and the fact that it was not explains the tragedy of those cases.[16]

Although the primary thesis of this paper is a defence of a modest mechanista position, I must admit that, on the question of what standard we should employ when evaluating historical cases like that of Semmelweis, the standard of consistency with actual (historical) scientific practice is unpersuasive. It is too big a bullet to bite to look back on cases like that of Semmelweis and maintain that the proper local principle of inference was the one that contributed to sustaining a false theory for many decades and which in turn caused the death of many thousands of new mothers. A tertiary guiding principle might be: no philosophical theory should have needless death on its hands.

The way to resolve this, I think, is to adjust the standard that mechanistas appeal to in assessing historical cases. They should, like the statisticalistas, employ a standard of prescriptive adequacy. After all, what is ultimately at stake in this debate is a prescriptive question: what kinds of evidence should be used when making causal inferences? To answer a prescriptive question while drawing on historical cases we should employ prescriptive standards of adequacy (though of course, if one's prescriptions were wildly at odds with scientific practice that would be a problem). Between is and ought we should mind the gap.

Two final considerations. First, since the above analysis relies on subjective Bayesianism, one might worry that the analysis inherits standard problems associated with subjective Bayesianism. Perhaps the most important criticism of Bayesianism is the 'problem of the priors', which holds that there is little constraint on the value that priors can take. For a given hypothesis H, one scientist could have her value for P(H) while another scientist could have a different value. This criticism is often exaggerated, since subjective Bayesianism is not the view that the pertinent probabilities are *merely* subjective; these probabilities are constrained by empirical facts and theoretical considerations. One might further worry that such constraints do not determine precise values for the pertinent probabilities. But the above analysis of the statisticalista and mechanista arguments does not require precise values for the probabilities. The above analysis relies only on the much coarser-grained requirements that the pertinent probabilities be extremal or middling. In many cases empirical facts and theoretical considerations can provide justification and intersubjective constraint for at least that much.

Second, the relative importance of statistical to mechanistic evidence assumed by many medical scientists and regulators is poised to flip if the movement known as 'personalised medicine' achieves more prominence. At present, as noted above, mainstream medical science and the regulation of medical products is committed to the statisticalista position. But personalised medicine is thoroughly committed to the mechanista position. The fundamental basis of personalised medicine is to develop finer-grained disease categories based on mechanistic reasoning, and to employ interventions that are tailored to these more specific disease categories, where such tailoring is also based on mechanistic reasoning. Indeed, some proponents of personalised medicine are calling for radical changes to current regulatory regimes such that randomised trials would no longer be required for approving new drugs.[17] Thus, many proponents of personalised medicine are calling for a switch from the radical statisticalista status quo to a radical mechanista alternative. This mechanista alternative would be more radical than the position argued for here, insofar as the modest mechanista position argued for here does not entail that we should dispel with randomised trials in standard cases.

6. Conclusion

This paper offers a new way of assessing a fundamental debate between two polarised views regarding the kinds of evidence that are required for causal inference regarding interventions in applied contexts such as medicine ands social policy. The two opposed positions in this debate are each warranted by two key arguments. Appealing to Bayesianism allows me to articulate the conditions under which the four arguments are persuasive.

This approach provides some vindication to a modest mechanista thesis. The Bayesian tells us to base our inferences on both the likelihood and the prior, and the prior is determined, at least in part, by appealing to relevant theoretical and mechanistic knowledge. Moreover, knowing whether or not the statisticalista arguments apply in any particular case requires appealing to mechanistic knowledge or the absence of it. The mechanista requirement that causal inferences in medicine and the social sciences be based on both statistical and mechanistic evidence is vindicated by our best general theory of inference. Though the arguments under investigation here have been assessed by a number of previous articles, the approach taken here offers a novel lens to view the relevant arguments. Since the approach taken here has been to assess two local principles of scientific inference by appealing to a more general, more fundamental, and independently justified theory of scientific inference, the approach taken here could be seen as offering some resolution to this debate.

[15] See, for example, (Russo, 2007), (Howick, 2011), (Broadbent, 2011), (Clarke, 2014), and diving further into historical details, (Scholl, 2013).

[16] The Semmelweis case has been a staple among philosophers as an illustration of scientific reasoning more generally. See (Hempel, 1966), (Lipton, 1991), (Bird, 2010), (Gillies, 2005), and (Tulodziecki, 2013).

[17] Such calls come from industry, medical scientists, and even some philosophers; see, for example, (Teira, 2017) and (Tonelli & Shirts, 2017). Though to be clear, most arguing for this position do so on the grounds that clinical trials are impossible for fine-grained disease categories with very few patients.

Acknowledgements

I am grateful to Mark Tonelli, Tim Lewens, Samuli Reijula, and Norma Daykin and audiences in various venues for discussion and feedback.

References

Andersen, H. (2012). Mechanisms: What are they evidence for in evidence-based medicine? *Journal of Evaluation in Clinical Practice, 18*(5), 992–999.
Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanistic alternative. *Studies in History and Philosophy of the Biological and Biomedical Sciences, 36*, 421–441.
Biddle, J. (2007). Lessons from the Vioxx debacle: What the privatization of science can teach us about social epistemology. *Social Epistemology, 21*(1), 21–39.
Bird, A. (2010). Eliminative abduction: Examples from medicine. *Studies In History and Philosophy of Science Part A, 41*(4), 345–352.
Broadbent, A. (2011). Inferring causation in epidemiology: Mechanisms, black boxes, and contrasts. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 45–69). Oxford University Press.
Cartwright, N. (2012). Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*(5), 973–989.
Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*(2), 339–360.
Claveau, F. (2012). The Russo–Williamson heses in the social sciences: Causal inference drawing on two types of evidence. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(4), 806–813.
Department of Clinical Epidemiology and Biostatistics. (1981). How to read clinical journals: V: To distinguish useful from useless or even harmful therapy. *Canadian Medical Association Journal, 124*(9), 1156–1162.
Fuller, J., & Flores, L. J. (2015). The risk GP model: The standard model of prediction in medicine. *Studies in History and Philosophy Biological and Biomedical Sciences, 54*, 49–61.
Gillies, D. (2005). Hempelian and Kuhnian approaches in the philosophy of medicine: The Semmelweis case. *Studies in History and Philosophy Biological and Biomedical Sciences, 36*(1), 159–181.
Goldacre, B. (2012). *Bad pharma: How drug companies mislead doctors and harm patients*. HarperCollins.
Grüne-Yanoff, T. (2016). Why behavioural policy needs mechanistic evidence. *Economics and Philosophy, 32*, 463–483.
Hempel, C. (1966). *Philosophy of natural science* (Englewood Cliffs).
Howick, J. (2011a). *The philosophy of evidence-based medicine*. BMJ Books.
Howick, J. (2011b). Exposing the vanities - and a qualified defense - of mechanistic reasoning in health care decision making. *Philosophy of Science, 78*(5), 926–940.
Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach: La salle*.
Illari, P. (2011). Mechanistic evidence: Disambiguating the Russo-Williamson thesis. *International Studies in the Philosophy of Science, 25*(2), 139–157.
Jukola, S. (2017). On ideals of objectivity, judgments, and bias in medical research – a comment on Stegenga. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 62*, 35–41.
Leibovici, L. (2001). Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: Randomised controlled trial. *BMJ British Medical Journal, 323*(7327), 1450–1451.
Leuridan, B., & Weber, E. (2011). The IARC and mechanistic evidence. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 91–109). Oxford University Press.
Lipton, P. (1991). *Inference to the best explanation*. London: Routledge.
Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Marchionni, C., & Reijula, S. (2019). What is mechanistic evidence, and why do we need it for evidence-based policy? *Studies in History and Philosophy of Science, 73*, 54–63.
Parkkinen, Veli-Pekka, Wallmann, Christian, Wilde, Michael, Clarke, Brendan, Illari, Phyllis, Kelly, Michael P., Norell, Charles, Russo, Federica, Shaw, Beth, & Williamson, Jon (2018). *Evaluating evidence of mechanisms in medicine: Principles and procedures. Springer Briefs in philosophy*. Cham: Springer International Publishing.
Pinto, M. F. (2015). Commercialization and the limits of well-ordered science. *Perspectives on Science, 23*(2), 173–191.
Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*, 157–170.
Scholl, R. (2013). Causal inference, mechanisms, and the Semmelweis case. *Studies In History and Philosophy of Science Part A, 44*(1), 66–76.
Smith, G., & Pell, J. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ, 327*, 1459.
Sober, E. (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press.
Sprenger, J. (2014). Bayesianism vs. Frequentism in statistical inference. In A. Hájek, & C. Hitchcock (Eds.), *Oxford handbook on philosophy of probability*. Oxford University Press.
Steel, D. (2007). *Across the boundaries*. Oxford University Press.
Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences, 42*, 497–507, 2011.
Stegenga, J. (2015a). Effectiveness of medical interventions. *Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 34–44, 2015.
Stegenga, J. (2015b). Herding QATs: Quality assessment tools for evidence in medicine. In Silberstein Huneman, & Lambert (Eds.), *Classification, disease and evidence: New essays in philosophy of medicine*. Springer.
Stegenga, J. (2018). *Medical Nihilism*. Oxford: Oxford University Press.
Straus, S. E., Richardson, W. S., Glasziou, P. P., & Haynes, R. B. (2005). *Evidence-based medicine: How to practice and teach* (3rd ed.). London: Elsevier Churchill Livingstone.
Teira, D. (2017). Testing oncological treatments in the era of personalized medicine. In G. Boniolo, & M. Nathan (Eds.), *Philosophy of molecular medicine* (pp. 236–251). Routledge.
Tonelli, M., & Shirts, B. (2017). Knowledge for precision medicine: Mechanistic reasoning and methodological pluralism. *JAMA, 318*(17), 1649–1650.
Tulodziecki, D. (2013). Shattering the myth of Semmelweis. *Philosophy of Science, 80*(5), 1065–1075.
Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science, 32*(1), 33–61.