Establishing the effectiveness of medical treatments is one of the most important aspects of medical practice. Bradford Hill’s viewpoints play an important role in inferring causality in medicine, and EBM+ seeks to improve evidence-based medicine, which is influential in establishing treatment effectiveness. At EBM+’s foundations lies the Russo–Williamson thesis (RWT), which can be seen as providing a reduction of Hill’s viewpoints into those involving difference-making and mechanistic evidence, both of which are claimed by the RWT’s proponents to be typically required for establishing causal claims in medicine. Yet little has been written on whether and how the RWT, EBM+, and Hill’s viewpoints establish treatment effectiveness. This could be because of uncertainty over what treatment effectiveness is. I provide an account of treatment effectiveness, analysing the role of the RWT, EBM+, and Hill’s viewpoints in this regard. I argue that Hill’s viewpoints and EBM+ can be useful in helping to evaluate evidence, but cannot directly establish or confirm treatment effectiveness. This is partly because effectiveness, I claim, is subject to inductive risk and therefore determined by non-epistemic values, neither of which Hill’s viewpoints nor EBM+ offer guidance on. I conclude by reinterpreting Hill’s viewpoints in light of establishing treatment effectiveness.

Introduction

In contemporary medical research and practice, much of the evidence regarding the putative effectiveness of medical treatments is based on randomised controlled trials (RCTs) and an evidence-based medicine (EBM)–inspired epistemic framework. Although EBM has been defined as ‘the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients’ (Sackett et al. 1996), as Bluhm (2017a, 210) notes, ‘[t]he central idea underlying EBM is that research can be ranked hierarchically based on study design,’ with RCTs and their meta-analyses sitting at the apex, and research on physiological mechanisms at the bottom. There are limitations with this approach (Cartwright and Stegenga 2011;
Stegenga 2011; Reiss 2015), however, particularly for non-pharmacological treatments (NPTs), for which placebo-controlled RCTs may be unsuitable and difficult (if not impossible) to construct (Gupta 2007; Walker and Rogers 2014; Kirsch, Wampold, and Kelley 2015; Savulescu, Wartolowska, and Carr 2016; Jukola 2019). Accordingly, alternative approaches to evidence generation/evaluation, such as those that give more epistemic weight to mechanisms, may, as advocated by EBM+ (ebmplus.org), be better suited than EBM to evaluating the effectiveness of medical treatments.

EBM+ seeks to correct what it sees as some of EBM’s shortcomings, and consequently has brought further helpful attention to the importance of mechanisms in evaluating causal claims (Clarke et al. 2013, 2014; Dragulinescu 2017; Williamson 2019). The chief claim of EBM+ (Parkkinen et al. 2018, 3), ‘which adopts the explicit requirements of EBM, to (1) make all the key evidence explicit and (2) adopt explicit methods for evaluating that evidence,’ is that ‘[e]vidence of mechanisms should be integrated with evidence of correlation to better assess causal claims.’ A driving force behind EBM+ is the Russo–Williamson Thesis (RWT), which states that two things are typically needed to establish a causal claim in medicine: ‘first, that the putative cause and effect are appropriately correlated; second, that there is some mechanism which explains instances of the putative effect in terms of the putative cause and which can account for this correlation’ (Williamson 2019, 33).1 Difference-making evidence shows that the effect varies with the cause, whereas mechanistic evidence shows that a mechanism, such as a biochemical pathway, exists (Illari 2011). While both types of evidence can come from clinical trials,2 mechanistic evidence can also come from in vitro studies, animal experiments, and biomedical imaging (Parkkinen et al. 2018). Some benefits of mechanistic evidence are that it helps avoid confounding by tracing the causal process between the treatment and its effects, while difference-making evidence helps avoid masking, and thus the two support each other, overcoming each other’s weaknesses (Illari 2011). Nonetheless, the RWT is controversial. Philosophers have pointed out possible counterexamples (Howick 2011a; Claveau 2012, 812), scepticism regarding the joint necessity of difference-making and mechanistic evidence (Broadbent 2011; Howick 2011a; Reiss 2012; Fiorentino and Dammann 2015; Solomon 2015), conceptual worries (Howick 2011a), and critiques and clarifications of Russo and Williamson’s treatment of mechanisms (Dragulinescu 2012). Some authors have held that difference-making evidence is sufficient but not necessary for establishing a causal claim (Solomon 2015; Landes, Osimani, and Poellinger 2018, 41), or that ‘mechanisms are necessary but not sufficient for causation claims’ (Dragulinescu 2017, 47). Others have tried to strengthen the thesis by distinguishing different versions of it (Gillies 2019a) and resolving ambiguities (Illari 2011). These refinements have led to EBM+, further work on which can be valuable if it can clarify how it and the RWT can contribute to establishing or confirming treatment effectiveness.

The RWT concerns what makes for good evidence in medicine, and holds that mechanistic evidence should be viewed on an equal basis with that of difference-making evidence in evaluating the effectiveness of medical treatments (Russo and Williamson 2011a, 576; Parkkinen et al. 2018). Austin Bradford Hill’s viewpoints have also been used to evaluate evidence to help establish causal claims in epidemiology. Indeed, Hill’s viewpoints, in articulating mechanistic and difference-making considerations, can be seen to express key features of the RWT (Russo and Williamson 2007):
Williamson (2021) provides a framework to help determine which and how many of Hill’s viewpoints, when satisfied, can establish causality, and Luján and Todt (2020) characterise the RWT as a reduction of Hill’s viewpoints to considerations of difference-making and mechanistic evidence. However, neither the RWT, EBM+, nor Hill’s viewpoints—all of which are related to each other—are explicit as to how they can inform treatment effectiveness claims. Accordingly, my focus in this article is analysing whether, and if so, how, the RWT, EBM+, and Hill’s viewpoints contribute to establishing or confirming treatment effectiveness, and then suggesting ways to build upon these approaches.

Immediately, at least with respect to the RWT ostensibly not being meant for this purpose, I could be charged with engaging a strawman. Illari and Russo (2014, 53, italics in original), for example, state, ‘The point [of the RWT] is not to judge causal conclusions drawn by scientists either in the past or today, but to understand what evidential elements science uses in drawing causal conclusions.’ Yet the RWT and its associated epistemological framework do apparently aim to establish sufficiency conditions for establishing causal claims, despite the RWT’s avowed purpose of simply stating necessary conditions and the RWT’s distinction between evidence that establishes causation and evidence that merely warrants action (Williamson 2019, 36). For example, based on case studies advanced by its proponents, by ‘establish’ the RWT appears to also mean stating evidence sufficient to justify claims that could very well be considered a call to action for scientists, clinicians, or public policy makers (e.g. ‘prescribe this treatment, because it is effective’). Ostensibly then, the RWT can be viewed as serving not merely as a test of whether a putative causal claim (e.g. ‘a treatment is effective’) has been legitimately established, but as a method for establishing it in the first place. Luján and Todt (2020), for example, apply the RWT to regulatory science, using it as a hypothetical test of whether an already regulated substance (such as a treatment) should continue to be regulated, or an unregulated substance should be regulated. If the RWT did not somehow indicate what the sufficient conditions for establishing or confirming treatment effectiveness are it might appear hollow in the face of researchers’ and clinicians’ needs for actionable information. Parkinnen and colleagues’ (2018) framework of grading evidence indeed appears to provide a decision procedure by which the status of a causal claim can be assessed based on the quality of the evidence, such as done by the Grading of Recommendations, Assessment, Development and Evaluation (GRADE; Guyatt et al. 2008) approach, with GRADE being used to not only rule out claims but to rule them in in the first place.

What the sufficient conditions are for establishing or confirming treatment effectiveness may differ depending on the disease/condition being treated and the treatment in question. This appears to accord with Russo and Williamson (2011b, 63), who think that whether a plausible mechanism or confirmed mechanism is needed for a causal claim in medicine to be established ‘depends on the epistemic context.’ Since they also suggest (Russo and Williamson 2011b, 65) that purpose plays an important role in determining the contours of a causal epistemology, at heart this seems to countenance a view whereby sensitivity to actual medical practice is important to not only understand but to justify making effectiveness claims. Understanding how the RWT and EBM+ have been used to establish or confirm treatment effectiveness might thus help clarify the limits of the RWT and EBM+, improve understanding of treatment effectiveness itself, and elucidate their interrelationships. This is particularly important because what it means for a
treatment to be effective in the first place is unclear. Accordingly, I first canvas several approaches in medicine to establishing or confirming treatment effectiveness, before explicating my own approach. I then use this as a way of understanding what role the RWT and EBM+ can and should play in helping to establish or confirm a treatment as being effective. Before concluding, I reinterpret Hill’s viewpoints—which have been almost exclusively applied to disease aetiology—to make them more applicable for helping to establish or confirm treatment effectiveness.

### Standard Approaches to Treatment Effectiveness

There are few philosophical accounts of treatment effectiveness (some include Ashcroft [2002] and Stegenga [2015]). In medicine, EBM, through its hierarchies of methods of generating evidence, specifies the sorts of evidence that warrant a treatment as being considered effective, and how such claims should be established. One common method of ascertaining effectiveness—and what could be considered the standard EBM approach—is called simple extrapolation. This method of generalising RCT results involves transporting effect sizes from RCTs to a target population, unless there is a compelling reason not to (Post, de Beer, and Guyatt 2013). This approach, however, is problematic. Some problems include patients in clinical practice possibly being markedly different from the participants in RCTs, publication bias (meaning that the studies that are published may give a distorted view of a treatment’s effects in comparison with any unpublished studies), inattention to the mechanisms by which treatments work, lack of mathematical justification, and the fact that some individuals respond differently than others to the same intervention, even if they are all representative of the study population in clinically relevant aspects (Stegenga 2015; Fuller 2021; Tresker forthcoming). Moreover, regardless of the type of study, study results do not represent or automatically determine the effectiveness of the treatment studied. Rather, as I have argued elsewhere, they serve as evidence for inferences to effectiveness claims (Tresker forthcoming). This distinction may seem obvious but it can be easily overlooked when studies, especially RCTs, are viewed as the means and the standard by which the effectiveness of treatments is established. Yet an account of effectiveness that merely or mostly looks at how a treatment performs in a given trial may not accurately reflect how the treatment will perform in a specific individual in clinical practice.

Even if RCT results could be simply transferred to a patient in clinical practice, effect sizes still need to be put in a form comprehensible to patients to serve as the basis for a treatment decision. One common measure of effectiveness used to counsel patients and communicate among clinicians the benefit of a treatment is the number needed to treat (NNT). The NNT can be seen as exemplifying a method of utilising effect sizes since it is the reciprocal of the absolute risk reduction, and therefore is derived directly from an RCT’s effect size. The NNT is the number of patients that would be needed to be treated to avoid a given outcome, such as a heart attack. Prima facie, this appears like a useful metric to assess effectiveness and guide patients in making treatment decisions. However, the NNT is prone to being misunderstood, such as if it is not made clear to patients that it is not an objective measure of effectiveness but is instead specific to a particular comparison from an RCT. As Stang and colleagues (2010, 820) warn, ‘Without stating the direction of the effect, the alternative treatment, the treatment period, and
the follow-up period, information in terms of NNTs is uninterpretable.’ Even with this information, however, NNTs may provide poor measures of effectiveness for patients in clinical practice because NNTs cannot easily be used to compare treatment options. To validly do this the different treatments would need to have been tested in similar populations, diseases (including similar disease stages), and timeframes, and use the same outcomes and comparator treatments (McAlister 2008). In the absence of this, decisions based on NNTs could be misleading.

Other approaches that rely on RCTs to assess the potential benefits of treatments include effectiveness scales, such as that used by the National Comprehensive Cancer Network (2021):

(1) (Palliative only): Provides symptomatic benefit only
(2) (Minimally effective): No, or unknown impact on survival, but sometimes provides control of disease
(3) (Moderately effective): Modest impact on survival, but often provides control of disease
(4) (Very effective): Cure unlikely but sometimes provides long-term survival advantage
(5) (Highly effective): Cure likely and often provides long-term survival advantage

While this scale could be useful under certain circumstances, for medical conditions in which outcomes are markedly different other scales would be needed. One general shortcoming of effectiveness scales like this is that in addition to mostly relying on RCTs at the expense of other types of evidence, such scales require agreement among its users in terms of which outcomes are most representative of effectiveness. Yet such scales are typically created and used by regulatory authorities or other organisations without incorporating the patient’s voice or being able to be adjusted by a patient or clinician in clinical practice. They thus may offer a limited sense of effectiveness by not reflecting different standards of effectiveness.

Evidence interpreted in the light of tools such as effectiveness scales and the NNT may be easy to digest by policy makers, clinicians, or other stakeholders but offer rather blunt and sometimes fallacious inferences to—or (by omitting non-RCT evidence) impoverished views of—a treatment’s effectiveness for a patient in clinical practice. Reliance on unitary, simplistic measures of effectiveness is therefore unrealistic, potentially misleading, and frustrates a comprehensive inferential account needed for a treatment effectiveness claim. This is not to impugn their value, however. Such approaches can serve important roles in understanding and simplifying medical evidence. But they are fundamentally simplifications that do not account for the complexity and context of effectiveness claims and the purposes for which they exist. It is doubtful that a single heuristic or tool could capture the widely different situations inherent in what role effectiveness plays for individual patients in clinical practice. Even a complex epistemological framework like EBM is unsuited for this end, partly because of its reliance on hierarchies of methods of generating evidence (Stegenga 2014).

Effectiveness may nonetheless seem to have a unitary meaning, especially for drugs, where the background assumptions underwriting effectiveness attributions are implicitly understood by stakeholders. This can happen for effectiveness claims arising in the context of study publications of RCTs, whereby an EBM framework is implied. If a
doctor says that a Food and Drug Administration (FDA)—approved drug is effective for treating diabetes, it will be understood that—on the basis of FDA approval procedures—the drug has shown superiority to placebo in typically at least two RCTs, with an acceptable adverse event profile. Yet what this conception of ‘effectiveness’ means can be highly subject to factors irrelevant to or even misleading with respect to what a particular patient in clinical practice might need to know. For example, special-interest groups campaigned for the approval of flibanserin (a drug used to treat low sexual desire in women), despite the FDA previously having deemed it to have an unfavourable benefit–risk ratio. The evidence of effectiveness of this drug thus took on different meanings if evaluated in a medical or policy context—here the medical context reflecting a weighing of clinical benefits with clinical risks, and the policy context incorporating economic and normative considerations. This illustrates how the standards for evidence for influencing policy can shift depending on context (La Caze and Colyvan 2017, 11), with possible distorting effects on effectiveness claims for specific patients.

Treatment effectiveness claims are often conveyed via trialists’ statements in journal articles’ discussion and conclusions sections. Such effectiveness statements reflect unspoken norms of effectiveness attribution among members of the clinical and research community, regulatory guidance on what constitutes a result of clinical relevance, and/or the personal views of the articles’ authors. Such statements can be wrong, however. Relying on authors’ declarations of effectiveness in trials’ publications, or interpretations provided in secondary sources, to support effectiveness claims is therefore of dubious utility and can lead to faulty decisions if the standard of effectiveness used by any given patient is markedly different.

Many of the differences underlying various standards of effectiveness may derive from different assessments of the relative values of different clinical-trial outcomes. No clinical-trial outcome is universally applicable to all treatments and medical conditions; outcomes are domain-specific and depend on the goals, stakes, and features of the treatment in question. Although assessing whether an outcome has been obtained is usually relatively straightforward, diverse opinions exist as to which thresholds reflect clinical significance and which outcomes are important to patients in the first place (Nowinski, Miller, and Cella 2017; Dankers et al. 2021). For example, patients—who typically lie downstream of clinical trials—may only have their voices heard in retrospect, after trial results have already been reported and interpreted, and then only in an ancillary way, deciding whether to take a treatment that has already been deemed effective according to criteria and methods they had no input into and may not understand. If patients can only state their preferences ‘after-the-fact’ on predetermined outcomes, they are not being afforded the opportunity to more integrally contribute to effectiveness evaluations.

For at least these reasons, a better approach to effectiveness is needed.

**An Account of Treatment Effectiveness**

Effectiveness claims, I have previously suggested (Tresker forthcoming), are predictions based on arguments for how target patients/populations will fare after taking a treatment in a given set of circumstances, compared with the counterfactual of not taking that treatment. Effectiveness claims involve the specific situation of an individual (such as a patient...
in clinical practice deciding what treatment to take) or population (such as regarding an effectiveness claim meant to apply to a certain group of people). For example, the claim that dapagliflozin is effective for the treatment of type 2 diabetes could mean several things: a patient taking the drug for 6 weeks will likely have a substantial reduction in glycated haemoglobin levels; taking the drug for a year could do the same and decrease the patient’s risk of hospitalisation for heart failure; or any number of other predictions. On a population level it could mean that if certain types of patients (say, those with a glycated haemoglobin level of 7–10% and inadequate glycaemic control with metformin) take the drug for 24 weeks they can expect, say, a 0.7% decrease in glycated haemoglobin levels (as Bailey et al. [2010] report). What would make the effectiveness claim reasonable to accept in such cases is reason to think these standards are therapeutically relevant with respect to the target condition, plus evidence that the treatment yields improvement according to these standards.

Importantly, there is no way to prove the truth of an effectiveness claim’s prediction, since even a person taking a treatment and improving clinically does not vindicate the particular effectiveness claim for that treatment. If effectiveness claims could be vindicated by simply observing an individual’s response to a treatment, then there would be no need for RCTs or large observational studies. Trying a treatment out is subject to far too many biases to in general be a reliable means of establishing valid treatment effectiveness claims. A person getting better after taking a treatment deemed effective only vindicates the claim insofar as it can contribute as evidence to other treatment effectiveness claims for the same treatment, and then only when interpreted in light of the many biases present (such as those articulated by Lilienfeld et al. [2014]), whereby there could be multiple competing reasons besides the treatment itself for why a person taking the treatment could have improved. For example, their response could be due to factors independent of the treatment, such as regression to the mean or spontaneous fluctuations in their condition. RCTs are useful insofar as they can help eliminate these alternative explanations on a population level via the assumption that the average response of these factors in both the treatment and the comparison group is approximately equal.

The existence of population-level and individual-level effectiveness claims raises the possibility that the two could conflict. For example, as indicated by the ‘prevention paradox’ (Rose 2008), whereby, as John (2011, 250) relates, ‘the most effective public health strategies have little benefit for the vast majority of those they affect,’ treatments such as statins could be considered effective on a population level but ineffective for most individual patients in clinical practice. Rather than undermine my account, I suggest that this possibility represents the fact that: (1) patients, even with similar clinically relevant characteristics (e.g. age, the presence of certain diseases), respond to treatments differently, and (2) the goals of effectiveness determinations on a population versus individual level often differ and should not be expected to coincide. Even solely on an individual level, though, a clinician could deem a treatment differentially effective for different patients. This reflects the heterogeneity of treatment effect commonly encountered in medicine. What then of different clinicians (ceteris paribus) deeming a treatment differentially effective for the same patient? Instead of rendering my account of effectiveness unacceptably fickle, this represents the common occurrence in medicine of different clinicians arriving at different conclusions (and why patients sometimes seek ‘a second opinion’). This does not mean the different clinicians are all correct about the treatment’s
effectiveness. Some effectiveness claims can be wrong. Accordingly, understanding why clinicians reach different conclusions offers opportunities for elucidating the reasoning and argumentation that are important for justifying treatment effectiveness claims.

A good account of treatment effectiveness, I stipulate, should possess at least some of the following features: effectiveness determinations should be predictive; can benefit from, but need not always require, a plurality of evidence; should ideally reflect the relevant stakeholders’ standard of effectiveness; and should facilitate the formulation of effectiveness claims that can be easily communicated and relied upon for treatment decisions. That effectiveness on my account inherently represents a prediction based on available evidence predisposes to being responsive to a plurality of evidence (including mechanistic evidence) and not simply the outcome of a clinical trial. By not privileging the results of RCTs as the arbiter of treatment effectiveness, my account of effectiveness as a prediction does not encounter some of the problems (as mentioned in the previous section) that beset the EBM approach that simply transports an effect size from a clinical trial to a patient in clinical practice.

An RCT’s outcome measure, indeed, may be inappropriate as the standard of effectiveness for a patient in clinical practice (Tresker forthcoming). Patients themselves could help choose a standard of effectiveness consistent with their values, preferences, and goals. Effectiveness on my account is responsive to which outcomes patients consider to be indicative of effectiveness because such outcomes can be the patient-informed standard of effectiveness for the treatment. Effectiveness claims on my account can therefore possibly be easily communicated to patients in terms they understand and can take action on. My account consequently considers important epistemic constraints of explanations of treatment effectiveness claims, exemplified by challenges some people have with numeracy when interpreting health information (Peters et al. 2007), the limited ability of some people to understand personalised medical risk calculators (Damman et al. 2017), and the fact that absolute and relative risks are perceived differently, by both doctors and patients (Perneger and Agoritsas 2011). Since a treatment effectiveness claim can involve a qualitative prediction (such as, ‘Will this treatment make me feel better?’), this potentially makes it easier than the EBM approach for patients to understand treatment effectiveness (though it should not downplay the usefulness of risk calculators and other ways of communicating risk to patients). The importance of this will hopefully become apparent in the context of inductive risk, which I discuss next.

The argument from inductive risk, early on developed by Rudner (1953) and Hempel (1965), and more recently by Douglas (2000, 2009), claims that non-epistemic values ‘help determine how serious it would be to make a false positive or false negative error, and thus how much evidence should be demanded in order to accept a hypothesis’ (Elliott and Richards 2017, 262). Inductive risk has mostly been discussed in the context of climatology and toxicology but, with the exception of its relationship to drug regulation and trial design (e.g. Plutynski 2017; Stegenga 2017; Bluhm 2017b), remains unexamined in the context of an account of treatment effectiveness. As considerations of science–policy communication differ among these fields, it is important to understand how inductive risk applies to treatment effectiveness, particularly regarding individual-level effectiveness claims.

A common belief in the inductive risk literature, as nicely summarised by Resnik (2017, 64), is that ‘scientists are morally obligated to consider the ethical or social
consequences of accepting or rejecting hypotheses when they are asked to provide advice to policymakers and the public. I, therefore, suggest that in the context of a clinical encounter patients fulfill the role analogous to the ‘policymaker’ and clinicians fulfill the role analogous to the ‘scientist.’ Treatment effectiveness claims, in my view, are not decisions by proxy as to whether patients should take treatments, but rather predictions as to what effects treatments might have if specific patients do take them. While it could be argued that the role of the clinician is to simply present information to patients about treatments, such as the results of clinical trials and how treatments work, patients also typically want to know whether a treatment will achieve a certain outcome for them—i.e. patients seek predictions to inform treatment decision-making.

The argument from inductive risk applied to treatment effectiveness could be viewed as the idea that non-epistemic (social, ethical) values determine how much and what type of evidence is needed for treatment effectiveness claims, which depend on the non-epistemic consequences of the risks of error—as Rudner (1953, 2) put it, ‘how serious a mistake would be.’ A treatment that were it to be deemed effective when really not (i.e. a false positive) might, for example, have serious consequences, such as patients being likely to take that treatment to the exclusion of other safer or less expensive alternatives. At the same time, the risk of false positives should be balanced against the risk of false negatives (missing out on calling a treatment effective that really is). To take an example I discuss briefly in the next section, art therapy could be deemed either effective or ineffective for the same patient depending on what the consequences of receiving art therapy would be were it in fact (in)effective. For example, assume the risks of adverse effects are low from a particular type of art therapy for a patient seeking to diminish their anxiety. Further assume the therapy is inexpensive, not inconvenient to the patient, and that there are no potentially more effective treatments available that would be foregone by engaging in the art therapy. The consequences of a false positive are therefore low—if the treatment doesn’t work not much is lost. The patient’s clinician, however, is faced with conflicting evidence on this art therapy’s effectiveness—while some RCTs indicate benefit, others do not, and the controls used in all the studies consisted of usual care or a no treatment wait-list (limitations of which are well known; see Tresker forthcoming). The clinician nonetheless determines, based also on mechanistic rationale, that if this patient engages in the therapy they will be likely to experience a moderate degree of diminished anxiety in comparison with the counterfactual of not engaging in the therapy. Alternatively, the clinician could make the less certain but equally warranted claim that the patient might experience the same outcome. The clinician could justifiably anticipate that the ‘likely’ claim will result in the patient electing to receive the treatment, whereas the ‘might’ claim will result in the patient refusing the treatment. There may be no way under the circumstances that the clinician can abdicate making any treatment effectiveness claim at all, because even if the clinician says nothing when asked if the treatment is effective, their reticence could still be interpreted by the patient as a ‘no’ or ‘not likely.’ The clinician making the more certain claim does so on the basis that the consequences of conveying that claim would be minor, a decision clearly influenced by non-epistemic values. Treatment effectiveness claims are therefore subject to inductive risk.

There are various points along the evidence chain for hypotheses (such as treatment effectiveness claims) at which inductive risk decisions are present. These include which
level of statistical significance to choose, how evidence is characterised, and how data are interpreted (Douglas 2000); the choice of model organism (Wilholt 2009); and numerous methodological aspects of trials, such as the choice of comparator (placebo or active comparator), outcome measure, and trial length (Plutynski 2017; Stegenga 2017; Bluhm 2017b). Deciding on the evidence sufficient to establish a standard of effectiveness as being met is also subject to inductive risk, a point made starkly clear by Stegenga (2017) in discussing the effectiveness of antidepressants in the face of significant publication bias. In summary, that non-epistemic concerns play a defining role in determining whether effectiveness is established for a particular treatment can be illustrated by the fact that inductive risk is involved in: (1) some of the decisions made by researchers in generating evidence on the effectiveness of treatments, even before a stakeholder (such as a clinician) has to decide whether to accept or reject a treatment effectiveness claim, (2) the decision on the standard of effectiveness, (3) the decision on what evidence and how much evidence is sufficient to establish that standard as being met, and (4) the decision on how the effectiveness claim should be conveyed (in terms of its formulation and any of its nonverbal features, including the clinician’s demeanour and aspects of the environment in which it is conveyed that might affect how the claim is interpreted).

One potential objection to applying the argument from inductive risk to treatment effectiveness could be that when researchers, clinicians, regulatory authorities, and other stakeholders accept or reject a hypothesis about whether a treatment is effective they need not do so tout court; instead, they can make the less-certain claim that a treatment is, for example, hardly/somewhat/moderately/very/extremely effective, and qualify any claims of effectiveness they make with varying degrees of probability—possibly, probably, definitely, etc. Indeed, as Steele (2012) points out, scientists’ beliefs about uncertainty often need to be converted into predefined evidence scales (a conversion which at least implicitly involves value judgments). Conveying degrees and likelihods of effectiveness are thus important communicative devices, reflections of the spectrum of effectiveness upon which treatments fall, and emblematic of the nuances captured by #4 of the previous paragraph. Thus, even if a hypothesis involves ‘hedged’ effectiveness claims it is still subject to inductive risk.

Because of the effects a communicated effectiveness claim can have on its recipients, such as licensing action (taking a treatment, regulating it, reimbursing it, etc.), I suggest that clinicians conveying treatment effectiveness claims to patients do more than merely accept hypotheses. Instead, treatment effectiveness claims should be viewed as assertions, a position that should not be too difficult to accept given that others (e.g. John 2015, 2019; Franco 2017) even hold this to be true of non–treatment-related hypotheses in general. Assertions better reflect the social and pragmatic dimensions of scientific practice (Franco 2017, 167–168). Accordingly, an important role for clinicians is thinking through how the treatment effectiveness claims they convey to patients will affect those patients.

Closely apposing effectiveness claims with likely actions (e.g. taking treatments) might encourage stakeholders to be more circumspect in their effectiveness claims and to be more explicit in articulating which non-epistemic values inform such claims. Douglas (2009, 155) suggests:
Scientists should be clear about why they make the judgments they do, why they find evidence sufficiently convincing or not, and whether the reasons are based on perceived flaws in the evidence or concerns about the consequences of error.

As noted by Elliott and Richards (2017, 272), ‘[i]t is impractical, however, to think that scientists could elucidate all the scientific judgments associated with a particular line of inquiry and clarify the roles that values played in making each of them.’ In the strict time limitations of a typical clinic visit, clinicians may only be able to convey to patients the most salient value considerations that influenced their inductive risk decisions in formulating a treatment effectiveness claim. Despite the infeasibility of such analyses and their disclosure to patients in clinical practice, such analyses could be valuable additions to medical education via case studies of individual-patient treatment effectiveness claims. On a population level, making explicit and defending the values involved in a treatment effectiveness claim could involve substantive argumentation that could be conveyed through narrative review articles or specialised investigations by regulatory authorities or other stakeholders. Such work could adduce a wide range of evidence in the context of an overriding argument in the specific situation in which the effectiveness determination needs to be made. Accordingly, I have refrained in this article from describing what is required for a treatment to be considered effective, and instead highlighted general constraints on effectiveness determinations. Towards this end, the close interconnection of effectiveness claims with how they are conveyed suggests that treatment effectiveness claims should be seen as part research and clinical science, and part medical communications. Plutynski (2017, 165) nicely summarises the issues at stake (with respect to screening, but also applicable to treatments): ‘At issue here then are fundamentally philosophical disagreements about justice, harm, autonomy, and beneficence, and the role of the physician with respect to both individual patients and the patient population more generally.’ Accordingly, the ways in which treatment effectiveness claims raise questions of harm and of paternalism (e.g. with respect to how clinicians modify their inductive risk calculus in relation to their presumptions about their patients’ psychology and anticipated decisions) offer an opportunity for the exposition of such cases to expand beyond the ambit of medical expertise to also involve patients as well as experts in non-medical fields, including philosophy, sociology, psychology, and economics.

**How do the RWT and EBM+ Establish Effectiveness?**

To provide an example of the RWT/EBM+’s approach to establishing treatment effectiveness I next discuss Gillies’s (2019a) analysis of the effectiveness of acupuncture. Relying on a single, secondary source from approximately two decades ago (Kaptchuk 2002), Gillies (2019a, 174–175) concludes: ‘It seems that RCTs definitely show that acupuncture is effective for adult postoperative and chemotherapy nausea, and for acute dental pain.’ This, for Gillies, provides the difference-making evidence. The mechanistic evidence he adduces not from Traditional Chinese Medicine (TCM), which he finds speculative and unscientific, but from contemporary scientific medicine. I concur with Gillies that plausible mechanisms exist for potential pain-relieving effects of acupuncture treatment. However, he does not cite or discuss any evidence that these exist for nausea. Moreover, to say that enough (and presumably of high-enough quality) difference-making evidence
exists for the effectiveness of acupuncture to be established (for the mentioned indications or any others) is perhaps premature, because it neglects discussion of the appropriate standard of effectiveness for acupuncture and a critical analysis of the available studies. For example, regarding acupuncture for postoperative and chemotherapy nausea, Kaptchuk (2002) cites a 1999 Cochrane review meta-analysis by Lee et al. which included 19 studies, 10 of which involved a no-treatment control group. This raises the question as to whether expectations, the treatment ritual, mere participation in a clinical trial, and/or response bias, as opposed to any effect of the needling, accounted for the clinical benefit. If perhaps Gillies means for the acupuncture treatment package (i.e. needling delivered by an empathetic practitioner with a plausible rationale and other trappings of a healing clinical environment) to be considered effective and not acupuncture itself (e.g. needling delivered to a person with postoperative nausea somehow unaware they are being given the treatment) then it might be fair to say that one of these treatments is effective and the other is not. However, to claim that the treatment package is effective would require a standard of effectiveness for such types of treatment packages. For effectiveness, superior performance to a no-treatment control group is a low bar to meet.

Perhaps then an updated, more rigorously controlled meta-analysis (Lee and Fan 2009) can provide the difference-making evidence to substantiate Gillies’s assertion. This meta-analysis compared acupuncture in the TCM-sanctioned point to the illusion of this point or to a non-TCM-sanctioned point. There was a relative risk of 0.71 (95% confidence interval 0.61–0.83), with moderate heterogeneity ($I^2 = 60\%$). Whether this should be considered effective has been challenged (Colquhoun and Novella 2013). Also, both meta-analyses included modalities other than acupuncture, such as acupressure, oppugning how representative the results are of acupuncture per se. An even more-recent updated meta-analysis (Lee, Chan, and Fan 2015) included the authors’ conclusions as indicating there being low-quality evidence supporting the claim.

Since numerous studies have shown that the effects of acupuncture do not necessarily depend on the location of needle placement or whether they are inserted (reviewed in Dincer and Linde 2003), it is rebuttable whether the ‘acupuncture’ Gillies begins talking about (i.e. the one explained by TCM theory) can be considered the same treatment as ‘acupuncture’ explained according to contemporary scientific medicine. This is because the latter could involve the insertion of needles at non-TCM points, or the absence of insertion at all, whereas under TCM theory needles inserted at non-TCM points do not constitute acupuncture.

I am not saying that some form of acupuncture may not be effective for some conditions. I am saying, however, that too easily accepting the pronouncements of experts or the medical community (ignoring the existence of multiple communities, disputing voices, or countervailing rationales) can lead one to arrive at a simplistic, incorrect conclusion regarding effectiveness. Moreover, it can lead to omitting key considerations that are important in evaluating the effectiveness of medical treatments, such as the legitimacy of the placebo comparison; also, focusing only on population-level effectiveness risks neglecting individual-level effectiveness.

Nonetheless, the RWT/EBM+ can be useful for disclaiming some effectiveness claims. For treatments such as homeopathy and retroactive intercessory prayer, for example, the RWT can serve as a disqualifier because of the absence of a plausible mechanism or the
postulation of a highly implausible one, despite the presence of RCT difference-making evidence (Russo and Williamson 2011a). These are cases of what Jerkert (2015) calls ‘negative mechanistic reasoning.’ In fact, Jerkert (2015) criticises the Lee and Fan (2009) Cochrane review for not taking into account any mechanistic considerations. The correct application of the RWT in such cases, however, does not necessarily extend support of it or EBM+ to more nuanced cases. These nuanced cases do not simply require a disqualification of an effectiveness claim but instead require positive warrant supporting one. Cases in which intuitions may clash, such as acupuncture, illustrate how simply knowing that there is some putative evidence of difference-making and mechanisms is insufficient to establish treatment effectiveness. This is because the roles such evidence plays can widely vary and be interpreted in multiple ways. The RWT does not specify what type or how much mechanistic evidence is required for establishing causal claims in medicine. Does mechanistic evidence always need to involve biological mechanisms, or would solely psychological or social mechanistic evidence count? Regarding the amount of evidence needed, Williamson (2021) states that the RWT does not require establishing the details of a mechanism, while other explanations of the RWT (e.g. Clarke et al. 2014, 357; Parkkinen et al. 2018, 94) emphasise the quality of the evidence. In this regard, Machamer, Darden, and Craver’s (2000) influential distinction between mechanism sketches, mechanistic schemata, and full mechanistic models might be a useful way of distinguishing between the extent of needed mechanistic evidence. Still, a way to decide upon a choice would be needed. Adopting the RWT and EBM+ is not a way to solve how treatment effectiveness can be established, but might better be interpreted as a framework that emphasises the importance of difference-making, and, in particular, mechanistic evidence, for establishing treatment effectiveness.

An ongoing theme underlying the RWT and EBM+ is engaging with the medical community. For example, the charts used by EBM+ (Parkkinen et al. 2018) to determine the status of mechanistic claims span the gamut from ‘ruled out’ to ‘speculative’ to ‘established,’ and therefore are judgments based on expert opinion. Illari (2011, 155) also thinks that ‘a certain amount of consensus’ is required for assessing the strength of certain forms of mechanistic evidence. The reliance by the RWT and EBM+’s proponents (e.g. Clarke et al. 2014) on the International Agency for Research on Cancer and other sources in medicine, such as the ‘medical community’ (e.g. Gillies 2019a, 148), to underwrite their application of the RWT are ways in which they have engaged with the medical community and apparently why they have accepted these agencies’ claims as authoritative. Given the plethora of causal claims that could be made in medicine it would be unrealistic to specify criteria by which each could be validly satisfied, so EBM+’s reliance on community standards and expert judgment is sensible in this respect. It is important to not forget, however, the various ways in which a scientific community could be wrong or in epistemic flux. This is illustrated by the acupuncture example, as well as the one discussed next.

Take an NPT—art therapy—as an example of a treatment heretofore unanalysed by the RWT or EBM+. The existence of some difference-making evidence (Uttley et al. 2015) combined with what might be considered good mechanistic evidence (Czamanski-Cohen and Weihs 2016) presents a situation whereby art therapy would appear to be effective (on a population level) according to the RWT. However, a pragmatic RCT (the MATISSE trial) showed art therapy, for one indication at least, to not
be effective (Crawford et al. 2012). Which conclusion is to be believed? Because of space, I will not perform a comprehensive analysis of the effectiveness of art therapy or how it can or should be evaluated, but as judged by critical editorials (Wood 2013; Holttum and Huet 2014) on whether the MATISSE trial really showed art therapy to not be effective, a valid and comprehensive assessment of art therapy is far more complex and nuanced than anything the RWT and EBM+ appear capable of. For example, a thorough evaluation that goes beyond the RWT might look to the legitimacy of the placebo controls employed in the existing RCTs, and a deep investigation of all the biases inherent in the existing clinical trials, analysing such evidence in light of the existing mechanistic evidence and skilfully synthesised in a way that a good narrative review article does. It is uncertain that the Is your policy really evidence-based? tool offered by EBM+ (Parkkinen et al. 2018), is capable of this sort of analysis, even when combined with expert judgment and GRADE (as EBM+ [Parkkinen et al. 2018, 27] suggests could be done to assess a correlation claim). For example, neither EBM+ nor GRADE facilitates an in-depth analysis of the legitimacy of placebo groups in terms of whether they offer valid comparisons to the experimental treatment. This, however, is crucial considering how widely a treatment’s effectiveness can be underestimated or overestimated depending upon the comparison (Howick and Hoffmann 2018). GRADE is also beset with other limitations (Rehfuess and Akl 2013) that it would be helpful for EBM+ to clarify if they expect GRADE to be used alongside evaluation of mechanistic evidence. Moreover, as alluded to in the previous section, a richer understanding of art therapy’s effectiveness should consider how its effectiveness differs on a population versus individual level, and how it is affected by decisions made under inductive risk.

According to Holman (2019, 4379), EBM+ is an example of friction-free epistemology in the sense that it does not take economic forces closely into consideration when evaluating evidence; he writes, ‘What such an approach is not well-equipped to do is make any recommendations about what inferences should be made given the current state of evidence or to make any policy recommendations about how the medical community should reform its institutions or epistemological priorities.’ Although Gillies (2019b) responds to some of Holman’s criticisms of EBM+, regardless of how well EBM+ takes economic forces into consideration EBM+ is useful insofar as it helps evaluate evidence. However, to help establish effectiveness EBM+ should be viewed in a framework of inductive risk and supplemented with guidance on navigating and establishing appropriate standards of effectiveness.

Overall, using the RWT and EBM+ as a general means of establishing treatment effectiveness stretches them beyond what they were formulated for or are capable of. Cases (i.e. acupuncture) where some of their proponents reach incorrect conclusions about effectiveness claims that have already been putatively established illustrate this. Granted, EBM+ is positioned as an aid to judgment, not a replacement (Parkkinen et al. 2018, 7). EBM+ is also an improvement over EBM in that it abandons hierarchies of methods of generating evidence, and lends more evidential weight to mechanistic evidence. Nonetheless, it is not entirely clear which other of EBM’s commitments EBM+ accepts and rejects. Rejection of EBM’s hierarchies and increased consideration of mechanistic evidence, while welcome improvements, do not necessarily produce a system or means by which the effectiveness of treatments can be established or
confirmed. Like EBM, EBM+ still involves rule-based reasoning, supplemented with guidance for mechanistic evidence and how to integrate this with difference-making evidence. Essentially, EBM+ is akin to an evidence assessment tool and is thus better seen as a method of assessing evidence as opposed to determining or confirming whether a treatment’s effectiveness should be considered established. Accordingly, the tools it proposes should be evaluated empirically and assessed for inter-rater and inter-tool reliability, as well as their validity and usability in terms of the applicability of items and how items should be weighted.17

If EBM+ and the RWT cannot establish or confirm effectiveness, this raises the question whether Hill’s viewpoints can. I address this next.

**Reinterpretation of Hill’s Viewpoints for Treatment Effectiveness**

Russo and Williamson (2011a, 572) claim that their interpretation of causality ‘provides a conceptual framework to underpin causal assessment as advised by Bradford Hill.’ Hill’s nine viewpoints (which are best seen not as causal criteria but as considerations to inform evaluation of causal claims) remain influential in philosophy and medicine. For example, Reiss (2015), in his pragmatist theory of evidence, implicitly seems to utilise at least some of Hill’s viewpoints. Stegenga (2011, 497) writes (in comparison with meta-analysis), ‘—the plurality of reasoning strategies appealed to by the epidemiologist Sir Bradford Hill—is a superior strategy for assessing a large volume and diversity of evidence.’ Howick and colleagues (2009) offer a revision and categorisation of Hill’s viewpoints and then apply them to examples of putative causation when RCT evidence is unavailable. Application of Hill’s viewpoints to diverse problems in medicine by clinicians and researchers is also popular (e.g. Kotsovilis and Slim 2012; Armon 2018; Smith et al. 2018). Even the GRADE approach to evidence evaluation, which is firmly rooted in EBM and is often used to assess treatment effectiveness by relying on a hierarchy of methods of generating evidence and not a plurality of evidence as championed by Hill, has been argued as reflecting Hill’s viewpoints (Schünemann et al. 2011). Indeed, common approaches to causal inference in philosophy and medicine, including GRADE, have been shown to overlap with Hill’s viewpoints (Shimonovich et al. 2021).

Hill’s viewpoints were developed with disease aetiology in mind (Hill 1965) and have mainly been applied to the establishment of disease causation.18 Based on an informal search on PubMed and PhilPapers.org, they appear to have been only infrequently applied to medical treatment, and there to the effectiveness of drugs—less so (if at all as far as I can tell) to the effectiveness of NPTs. Hill’s viewpoints have also typically been applied to situations in which evidence from observational studies, but not RCTs, is available. How they apply to evaluating a treatment for effectiveness, therefore, does not appear to be well known. I address this lacuna by charitably reinterpreting Hill’s viewpoints to see how they can help evaluate evidence for treatment effectiveness claims. Because Hill’s viewpoints were not designed for treatment effectiveness, my approach is revisionary (and preliminary), with no pretence at capturing Hill’s original meanings. My descriptions of some viewpoints may appear applicable
to other viewpoints. This could reflect that there are no clear boundaries between some of the viewpoints.

**Strength of Association**

A strong association between a treatment and an outcome could, in combination with Hill’s other viewpoints, indicate causality. Williamson (2019) rightly notes that large effect sizes observed in observational studies combined with excellent evidence of mechanisms offer sufficient evidence for effectiveness to be established, even in the absence of RCTs. However, strong causal associations can also entail weak effects, and strong correlations can be spurious. Knowing the causal structure of a treatment can make a strong observational association between that treatment and an effect more plausible as an indicator of causality. Here mechanistic knowledge can be useful. An example would be a drug that suddenly shrunk a person’s tumour far beyond what could be expected by chance occurrence or the natural history of the disease (see also the ‘coherence’ and ‘plausibility’ viewpoints).

**Consistency**

According to Hill (1965), consistency involves consistent findings or replication across different locations, populations, and methods. While this could be a useful indicator of a treatment being effective, this could also reflect an artefact in how treatments are studied or administered. Placebo effects, for example, are commonly encountered with most treatments, such as via conditioned responses and by being secondary to an empathetic clinician who provides support, encouragement, and expectation that a treatment will work. A benefit of the consistency viewpoint might instead rest with how it brings attention to the importance of treatments being defined consistently (i.e. between study and clinical settings). For example, a treatment should be consistently described as containing, or not containing, whatever aspects of the treatment produces placebo effects, depending upon whether the treatment’s standard of effectiveness includes that it should or should not work by placebo effects. I discuss this further under the ‘plausibility’ viewpoint.

**Specificity**

Instead of meaning (as originally formulated by Hill) that an exposure causes only one disease, specificity could be reinterpreted to mean whether a treatment works only for a specific disease and not others. Some treatments for inborn errors of metabolism or rare diseases, for example, could be seen as exemplifying this viewpoint. However, there are many treatments, such as broad-spectrum antibiotics, that are effective for many conditions.

The viewpoint could also be reinterpreted in terms of causal specificity as explained by Woodward (2010), whereby specificity represents the degree to which causes exert fine-grained influence over their effects. However, a relative lack of causal specificity can actually be a boon to successful intervention, with some
treatments (such as HIV drugs) involving less specific causes being more highly valued clinically (Neal 2019).

How useful this viewpoint is to establishing or confirming treatment effectiveness therefore currently seems unclear.

**Temporality**

Since for medical treatment it is typically clear which comes first—the treatment or the effects—rather than serving any useful purpose for establishing or informing a claim of effectiveness, this viewpoint merely serves as the precondition for a causal attribution. To be useful, temporality can be reinterpreted as whether the time course of a treatment’s effects is consistent with mechanistic knowledge, or can be used to establish what the putative mechanisms might be. For example, in the EMPA-REG OUTCOME trial of empagliflozin treatment in patients with type 2 diabetes (Zinman et al. 2015), the benefit on cardiovascular mortality was evident as early as 3 months, whereas previous large trials of glucose-lowering drugs had not shown a reduction in cardiovascular events until about 10 years of treatment, which argues against glycaemic control as being a significant mechanism by which empagliflozin benefits the heart.

**Biological Gradient, or Dose-response**

This may apply to some drugs but for many does not. For example, there was no difference between empagliflozin 10 mg and 25 mg on outcome measures in EMPA-REG OUTCOME (Zinman et al. 2015). For some NPTs, such as surgery, this viewpoint is largely irrelevant. Similarly, for other NPTs, such as psychotherapy, the phenomenon of responsiveness, whereby the therapist and client adjust the number of sessions according to the client’s need (Stiles et al. 2008; Stiles 2009), makes it very difficult to determine a consistent dose-response relationship.

**Plausibility**

This viewpoint can be seen to reflect whether plausible mechanistic evidence exists. Although this has traditionally been considered to involve mechanisms explainable on a biological level, given the diverse levels on which treatments work, it could also involve psychological and social mechanisms.

In terms of my account of treatment effectiveness, mechanisms are important for establishing effectiveness because of how they can contribute to valid treatment effectiveness claims by supporting the argument for a prediction about a treatment’s effects; moreover, they can serve an important communicative function when clinicians explain treatments to patients, thus contributing to how a patient understands a treatment effectiveness claim.

When interpreting epidemiological studies,

Evidence of mechanisms can help rule in or out various explanations of a correlation. For example, it can help to determine the direction of causation, which variables are potential confounders, whether a treatment regime [sic] is likely to lead to performance bias, and whether measured variables are likely to exhibit temporal trends. (Parkkinen et al. 2018, 16)
Mechanisms can also facilitate the synthesis of diverse kinds of evidence, generalisation of study results, what outcome measures to use in clinical trials, and other trial-design decisions (Clarke et al. 2014; La Caze and Colyvan 2017; Aronson et al. 2018). As mentioned for the ‘consistency’ viewpoint, when stakeholders view certain mechanisms as being crucial to how a treatment works, mechanistic knowledge can aid in formulating an unambiguous description of the treatment—a description that is used to define that treatment with respect to how the standard of effectiveness for it can be met, such as by informing, for example, what the control group in an RCT should comprise.

The reason a treatment should be clearly and unambiguously characterised is because only then can tests of a treatment be made whose results can be validly used as evidence for effectiveness claims for that same treatment, and not slightly different ones. When ‘covert treatment substitutions’ happen, the helping or hindering factors that are present/absent in a trial are not present/absent in the same way in the clinical setting in which the treatment is administered. While this could be viewed as the treatment’s effectiveness being contextual, it can also be viewed as the description of a treatment in clinical practice subsuming contextual factors, but differing from those factors subsumed during the trial. An unambiguous description of a treatment helps eliminate alternative explanations for the treatment’s effects and helps inform the standard of effectiveness. It does this by indicating what alternative explanations are deemed ineligible for effectiveness. This could guide the appropriate control group in an RCT. For example, a good standard of effectiveness for a treatment might involve it not working through placebo effects if stakeholders deem those effects not integral to how the treatment should work.

**Coherence**

This viewpoint was used by Hill to mean lack of conflict with the natural history and biology of the disease. When such information is known, this can be useful for ruling out alternative explanations for a treatment’s effects. ‘Coherence’ thus underscores the importance of mechanistic knowledge of diseases’ natural histories.

**Experiment**

How do experiments offer evidence for treatment effectiveness? Unlike disease aetiology, for which conducting RCTs is usually unethical or infeasible, the effectiveness of treatments (particularly pharmacological treatments) can be and typically are evaluated by RCTs. This seems like it would obviate the need for evidence from observational studies and hence the need to apply Hill’s viewpoints. Yet observational studies play important roles in evaluating treatment effectiveness (Black 1996), particularly for NPTs. The ‘experiment’ viewpoint then, at least with respect to treatment effectiveness, should not be seen as something separate from the other viewpoints, as if it could clinch treatment effectiveness and therefore diminish or eliminate the need for evidence from observational studies. Indeed, responses to some treatments have been shown to differ depending on if studied in RCTs or observational studies (Naudat, Maria, and Falissard 2011), and such average treatment responses may also differ from what could be expected with an individual patient in clinical practice. For both observational studies and RCTs
the target population/individual in clinical practice is different than the study population. As discussed in the ‘An account of treatment effectiveness’ section, clinical trial results therefore need to be generalised.

A clinical trial (of any type; see note 2) I have proposed can be seen as being generalisable to the extent to which its results can contribute to valid inferences to treatment effectiveness claims (Tresker forthcoming). Although generalisability has traditionally been considered to be a reflection of a trial’s external validity in terms of how representative the trial subjects are to those in the target population, I have also argued that representativeness, though important, is not a suitable feature for determining how well a trial contributes to valid treatment effectiveness claims (Tresker forthcoming). RCTs, though often very helpful for obtaining a treatment effect in a specific population, hold no privileged position in my account of treatment effectiveness when it comes to determining treatment effectiveness for a specific patient in clinical practice. The ‘experiment’ viewpoint should therefore be viewed as underscoring one of the many ways by which inferences to treatment effectiveness claims can be made. Looking at effectiveness this way could result in some nonintuitive conclusions, though. For example, Wilde and Parkkinnen (2019) argue that when difference-making evidence in humans is absent, mechanism-based extrapolation from animal models (combined with appropriate mechanistic evidence) is sufficient to establish an exposure as a cause of disease. Applied to medical treatment, this could entail that the effectiveness of a treatment could be established in the absence of human studies, using extrapolated difference-making evidence from animal studies. The view of effectiveness I have articulated is consistent with this possibility.

**Analogy**

Analogy can be reinterpreted to apply to medical treatment by considering whether similar treatments work in similar patients. For example, drugs of the same class often produce similar effects. But because they can also produce very different effects this viewpoint should be applied with caution.

The foregoing reinterpretations preliminarily indicate that Hill’s viewpoints could be useful for helping to establish or confirm a treatment as being effective. Indeed, approximately 14 years before his President’s Address to the Section of Occupational Medicine of the Royal Society of Medicine in which he expanded upon and refined the viewpoints first promulgated by the Surgeon General’s Advisory Committee on Smoking and Health (US Public Health Service 1964), Hill had alighted upon some principles that could be used to assess the effectiveness of treatments. These included replication of the result in similar patients and, ‘whether the result was merely due to the natural history of the disease or in other words to the lapse of time, or whether it was due to some other factor which was necessarily associated with the therapeutic measure in question’ (Hill 1951, 278; quoting Pickering 1949, 231). Hill thus adumbrated the ‘coherence’ viewpoint, and in doing so identified an important feature for helping to establish the effectiveness of treatments. Nonetheless, for Hill’s viewpoints to be more useful for establishing or confirming treatment effectiveness, a tenth viewpoint that reflects the dependence of effectiveness claims on non-epistemic considerations might be helpful. However, since demarcating the specific criteria that can identify the legitimate uses of
values (and which ones) in science is unsettled (Holman and Wilholt 2022), demarcating such criteria for treatment effectiveness is not something I can attempt here. Similarly, although I have indicated that an application of Hill’s reinterpreted viewpoints might correctly judge an intervention to be effective, such an investigation is only a beginning and also exposes an unresolved tension in just how the RWT expresses Hill’s original viewpoints. Further work can help clarify the application of the reinterpreted viewpoints to unequivocally effective treatments and those whose effectiveness is contested, to ultimately determine the usefulness of Hill’s reinterpreted viewpoints to establishing or confirming treatment effectiveness.

Conclusions

In this article, I have articulated an account of treatment effectiveness and described how the RWT, EBM+, and Hill’s viewpoints contribute to understanding and establishing treatment effectiveness. The RWT, when not stretched beyond what it was formulated for, is useful by highlighting the importance of mechanistic evidence and its typically joint necessity with difference-making evidence for contributing to good treatment effectiveness predictions. How much and what type of evidence is required for establishing treatment effectiveness, however, is constrained by the idiosyncrasies of specific treatments and the purposes and contexts in which they are used. Accordingly, the RWT and EBM+ should be seen as pieces of a much larger puzzle for establishing or confirming treatment effectiveness. My reinterpretation of Hill’s viewpoints is a modest attempt at expanding the repertoire of tools for establishing or confirming treatment effectiveness, especially when applied with an understanding of how effectiveness claims are fundamentally infused with values, some of whose influence should be addressed so that they can be managed appropriately. Areas I have suggested that could build on EBM+ include a greater sensitivity to inductive risk in formulating and communicating treatment effectiveness claims, and better procedures by which the medical community can assess standards of effectiveness and what amount, quality, and combination of difference-making and mechanistic evidence are sufficient to establish treatment effectiveness. Consistent with these suggestions, I have described an account of treatment effectiveness that emphasises non-epistemic features of treatment effectiveness claims, the importance of defending such claims and subjecting them to normative scrutiny, the desirability of establishing a clinically relevant and (particularly in the case of individual patients) patient-endorsed standard of effectiveness, and the predictive and rhetorical aspects of treatment effectiveness claims.

Notes

1. According to Williamson (2019, 36) “appropriately correlated” just means probabilistically dependent conditional on potential confounders, where the probability distribution in question is relative to a specified population or reference class of individuals.
2. Clinical trials include, inter alia, RCTs; observational trials whereby the investigator observes outcomes without providing an intervention, such as case-control, cross-sectional, and cohort studies; case series; and N-of-1 trials.
3. Though as Leuridan and Weber (2011) note, examples attempting to show the RWT to be undermined could also be interpreted as calls for the associated claim of effectiveness to be amended.

4. I thank an anonymous reviewer for this journal for the lattermost point.

5. In the interest of space, I only discuss my own, in the next section.

6. One might nonetheless try to rely on the efficacy-effectiveness distinction to argue that RCTs (or, specifically, explanatory RCTs as opposed to pragmatic RCTs) are used to assess the efficacy of treatments whereas observational studies (and pragmatic RCTs) assess effectiveness. This distinction, however, is problematic and typically either just indicates the method by which a treatment is studied, or the population/setting in question (i.e. a study/research setting, or target population/clinical practice), but does not indicate anything intrinsic about a treatment’s effectiveness (Tresker forthcoming). In the next section I offer an account of effectiveness that, in the case of an individual patient in clinical practice, looks at how a treatment is likely to affect that patient. In this regard, both so-called efficacy and effectiveness trials can produce evidence of effectiveness, since in both cases inferences will need to be made regarding how the treatment will affect a given patient in clinical practice.

7. For reasons of space I restrict my discussion of treatment effectiveness claims to the individual case and leave a discussion of treatment effectiveness claims for populations (such as might be made by researchers, regulatory authorities, health insurers, etc.) to potential future work.

8. Nonetheless, N-of-1 trials, although primarily just suitable for chronic, symptomatic conditions with easily measurable clinical outcomes, could provide useful information about treatment effectiveness, though typically only for the patient subject to the trial (Lillie et al. 2011; Duan, Kravitz, and Schmid 2013).

9. While a patient’s treatment decision can be considered in the context of a benefit–cost analysis, there may be good reasons for an effectiveness claim to be seen as an input to this analysis and not as equivalent to the outcome. In work in progress, I articulate how better understanding and facilitating treatment decision-making in the context of a clinical encounter, particularly from a shared decision-making standpoint, can help to better understand the contours of how treatment effectiveness claims can be integrated into decision theory.

10. This is not to say that plausible mechanisms of the effects of acupuncture on nausea do not exist; see Chen et al. (2014), for example, for some possibilities, although some acupuncture researchers claim that the mechanisms remain undiscovered (Shi et al. 2019).

11. Although Gillies’s account of acupuncture is problematic this should not be seen as impugning EBM+ as a whole, since alternative EBM+-inspired analyses might reach different conclusions. Moreover, Gillies should not be faulted, because rather than provide a rigorous substantive analysis of the effectiveness of acupuncture his chief aim appears to have been to illustrate the importance of mechanisms for establishing causal claims in medicine, for which I think he is successful.

12. ‘NegC’ reasoning in his typology whereby meta-mechanistic reasoning is used to argue that there is no plausible mechanistic chain between the putative cause and effect.

13. Illustrated by clinical practice guidelines (CPGs) when some experts hold diametrically opposed viewpoints. For example, CPGs promulgated by the American Academy of Orthopaedic Surgeons (2013) strongly recommended against viscosupplementation for the treatment of osteoarthritis (OA) of the knee, despite statistically significant results versus placebo, based on what they saw as clinical insignificance for some patients. This was met with strong resistance by physicians and industry members (Peer Review & Public Comments and AAOS Responses 2012). As Lubowitz, Provencher, and Poehling (2014, 4, italics in original) opined, ‘In our opinion, it is not in the interests of all patients to recommend against a treatment that is of significant benefit for some patients, especially when that treatment is for a disease (knee OA) that is not preventable, and for which there is no cure.’ This example also illustrates how non-epistemic values can influence effectiveness claims, and the importance of not conflating effectiveness determinations with treatment recommendations or decisions.
14. EBM+ proponents might argue that both sets of evidence should be evaluated and a conclusion drawn from that. However, it is unclear how such conclusions can be drawn using an EBM+ framework. Parkinnen and colleagues (2018, 74), for example, in offering an example of an intervention to which an EBM+ analysis could apply, subsequently state, ‘We did not undertake a systematic review of the evidence on how probiotics might work.’ This implies an actual substantive investigation is needed, possibly rendering the philosophical work otiose.

15. In this regard, use of the Template for Intervention Description and Replication (TIDieR-Placebo) checklist (Howick et al. 2020) could be used alongside EBM+’s checklists, given that it allows key features of both active interventions and placebo or sham controls to be concisely summarised by researchers, and also whether blinding was successful. Since this can easily be integrated with EBM, this seems like something that could also be integrated with EBM+.

16. As Howick (2011b) indicates, EBM’s view of evidence is fundamentally based on hierarchies of methods of generating evidence. If this is true then it is not clear how EBM+ merits the EBM appellation given its apparent abandonment of evidence hierarchies by viewing mechanistic evidence on the same footing as difference-making evidence.

17. Some of these may very well be concerns shared by members of the EBM+ community, so my arguments should be construed less as criticism and more as spotlighting areas where EBM+ can be built upon.

18. Despite their popularity, Hill’s viewpoints—even with respect to disease causation—have received their fair share of criticism (e.g. Rothman and Greenland [2005]). Hill’s viewpoints are not well-defined conditions and it is not clear what is required for any given viewpoint to be satisfied, much less how they can be weighted and integrated to establish causation (Phillips and Goodman 2006). According to Ward (2009), Hill’s viewpoints do not justify inferring that a statistical association is causal.

19. For example, Dammann (2018, 3) raises the point that since Hill didn’t require all of the viewpoints be fulfilled for a causal claim to be supported and no single one (with the exception of temporality) to be necessary, claiming as Russo and Williamson do that Hill’s guidelines express principles of the RWT elides over the fact that the RWT requires at least some of the viewpoints (namely, those expressing mechanistic and difference-making evidence) to be necessary.

Acknowledgements

Thanks to two anonymous reviewers for International Studies in the Philosophy of Science for helpful comments, and especially one of these reviewers for encouraging me to expand my account of treatment effectiveness, and for many other helpful comments. I am grateful to Federica Russo for reviewing an early draft and to Bennett Holman for reviewing one part of the final draft.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The author was supported by the Fonds Voor Wetenschappelijk Onderzoek – Vlaanderen (FWO) during the writing of this article (1130819N).

ORCID

Steven Tresker http://orcid.org/0000-0002-6511-4792
References

American Academy of Orthopaedic Surgeons. 2013. Treatment of Osteoarthritis of the Knee. 2nd ed. Evidence-Based Clinical Practice Guideline. https://www.aaos.org/globalassets/quality-and-practice-resources/osteoarthritis-of-the-knee/osteoarthritis-of-the-knee-2nd-edition-clinical-practice-guideline.pdf. Published May 18, 2013.

Armon, C. 2018. “From Snow to Hill to ALS: An Epidemiological Odyssey in Search of ALS Causation.” Journal of the Neurological Sciences 391: 134–140. doi:10.1016/j.jns.2018.05.016.

Aronson, J. K., A. La Caze, M. P. Kelly, V.-P. Parkkinen, and J. Williamson. 2018. “The Use of Mechanistic Evidence in Drug Approval.” Journal of Evaluation in Clinical Practice 24 (5): 1166–1176. doi:10.1111/jep.12960.

Ashcroft, R. 2002. “What is Clinical Effectiveness?” Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 33 (2): 219–233. doi:10.1016/S0039-3681(02)00020-1.

Bailey, C. J., J. L. Gross, A. Pieters, A. Bastien, and J. F. List. 2010. “Effect of Dapagliflozin in Patients With Type 2 Diabetes Who Have Inadequate Glycaemic Control With Metformin: A Randomised, Double-Blind, Placebo-Controlled Trial.” The Lancet 375 (9733): 2223–2233. doi:10.1016/S0140-6736(10)60407-2.

Black, N. 1996. “Why We Need Observational Studies to Evaluate the Effectiveness of Health Care.” BMJ: British Medical Journal 312 (7040): 1215–1218.

Bluhm, R. 2017a. “The Hierarchy of Evidence, Meta-Analysis, and Systematic Review.” In The Routledge Companion to Philosophy of Medicine, edited by M. Solomon, J. R. Simon, and H. Kincaid. Routledge. doi:10.4324/9781315720739

Bluhm, R. 2017b. “Inductive Risk and the Role of Values in Clinical Trials.” In Exploring Inductive Risk: Case Studies of Values in Science, edited by K. C. Elliott and T. Richards. Oxford University Press. doi:10.1093/acprof:oso/9780190467715.003.0010

Broadbent, A. 2011. “Inferring Causation in Epidemiology: Mechanisms, Black Boxes, and Contrasts.” In Causality in the Sciences, edited by P. M. Illari, F. Russo, and J. Williamson. Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199574131.001.0001/acprof-9780199574131-chapter-3

Cartwright, N., and J. Stegenga. 2011. “A Theory of Evidence for Evidence-Based Policy.” In Evidence, Inference and Enquiry, edited by P. Dawid, W. Twining, and M. Vasilaki, Vol. 71, 291–322. Oxford University Press. http://ukcatalogue.oup.com/product/9780197264843.do

Chen, S., S. Wang, P. Rong, J. Wang, L. Qiao, X. Feng, J. Liu, and J. Zhang. 2014. “Acupuncture for Visceral Pain: Neural Substrates and Potential Mechanisms.” Evidence-Based Complementary and Alternative Medicine, 609594. doi:10.1155/2014/609594.

Clarke, B., D. Gillies, P. Illari, F. Russo, and J. Williamson. 2013. “The Evidence That Evidence-Based Medicine Omits.” Preventive Medicine 57 (6): 745–747. doi:10.1016/j.ypmed.2012.10.020.

Clarke, B., D. Gillies, P. Illari, F. Russo, and J. Williamson. 2014. “Mechanisms and the Evidence Hierarchy.” Topoi 33 (2): 339–360. doi:10.1007/s11245-013-9220-9.

Claveau, F. 2012. “The Russo–Williamson Theses in the Social Sciences: Causal Inference Drawing on Two Types of Evidence.” Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43 (4): 806–813. doi:10.1016/j.shpsc.2012.05.004.

Colquhoun, D., and S. P. Novella. 2013. “Acupuncture Is Theatrical Placebo.” Anesthesia & Analgesia 116 (6): 1360–1363. doi:10.1213/ANE.0b013e31828f2d5e.

Crawford, M. J., H. Killaspy, T. R. Barnes, B. Barrett, S. Byford, K. Clayton, J. Dinsmore, et al. 2012. “Group Art Therapy as an Adjunctive Treatment for People With Schizophrenia: A Randomised Controlled Trial (MATISSE).” Health Technology Assessment (Winchester, England) 16 (8, iii–iv): 1–76. doi:10.3310/hta16080.

Czamanski-Cohen, J., and K. L. Wehls. 2016. “The Bodymind Model: A Platform for Studying the Mechanisms of Change Induced by Art Therapy.” The Arts in Psychotherapy 51: 63–71. doi:10.1016/j.aip.2016.08.006.
Dammann, O. C., N. M. M. Bogaerts, M. J. van den Haak, and D. R. M. Timmermans. 2017. “How Lay People Understand and Make Sense of Personalized Disease Risk Information.” *Health Expectations* 20 (5): 973–983. doi:10.1111/hex.12538.

Dammann, O. 2018. “Hill’s Heuristics and Explanatory Coherentism in Epidemiology.” *American Journal of Epidemiology* 187 (1): 1–6. doi:10.1093/aje/kwx216.

Dankers, M., M. H. J. M. G. Nelissen-Vrancken, B. H. Hart, A. C. Lambooij, L. van Dijk, and A. K. Mantel-Teeuwisse. 2021. “Alignment Between Outcomes and Minimal Clinically Important Differences in the Dutch Type 2 Diabetes Mellitus Guideline and Healthcare Professionals’ Preferences.” *Pharmacology Research & Perspectives* 9 (3): e00750. doi:10.1002/prp2.750.

Dincer, F., and K. Linde. 2003. “Sham Interventions in Randomized Clinical Trials of Acupuncture – a Review.” *Complementary Therapies in Medicine* 11 (4): 235–242. doi:10.1016/s0965-2299(03)00124-9.

Douglas, H. 2000. “Inductive Risk and Values in Science.” *Philosophy of Science* 67 (4): 559–579. http://www.jstor.org/stable/188707.

Douglas, H. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press. doi:10.2307/j.ctt6wrc78

Dragulinescu, S. 2012. “On ‘Stabilising’ Medical Mechanisms, Truth-Makers and Epistemic Causality: A Critique to Williamvos and Russo’s Approach.” *Synthese* 187 (2): 785–800. doi:10.1007/s11229-011-0011-9.

Dragulinescu, S. 2017. “Mechanisms and Difference-Making.” *Acta Analytica* 32 (1): 29–54. doi:10.1007/s12136-016-0292-1.

Duan, N., R. L. Kravitz, and C. H. Schmid. 2013. “Single-Patient (N-of-1) Trials: A Pragmatic Clinical Decision Methodology for Patient-Centered Comparative Effectiveness Research.” *Journal of Clinical Epidemiology* 66 (8 Suppl): S21–S28. doi:10.1016/j.jclinepi.2013.04.006.

Elliott, K. C., and T. Richards. 2017. “Exploring Inductive Risk.” In *Exploring Inductive Risk: Case Studies of Values in Science*, edited by K. C. Elliott and T. Richards. Oxford University Press. doi:10.1093/acprof:oso/9780190467715.003.0013

Fiorentino, A. R., and O. Dammann. 2015. “Evidence, Illness, and Causation: An Epidemiological Perspective on the Russo–Williamson Thesis.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 54: 1–9. doi:10.1016/j.shpsc.2015.09.010.

Franco, P. L. 2017. “Assertion, Nonempistemic Values, and Scientific Practice.” *Philosophy of Science* 84 (1): 160–180. doi:10.1086/688939.

Fuller, J. 2021. “The Myth and Fallacy of Simple Extrapolation in Medicine.” *Synthese* 198: 2919–2939. doi:10.1007/s11229-019-02255-0.

Gillies, D. 2019a. *Causality, Probability, and Medicine*. 1st ed. Routledge. doi:10.4324/9781315735542

Gillies, D. 2019b. A Reply to Holman’s Criticisms of EBM+. https://www.researchgate.net/publication/337544431_Holman’s_Criticisms_of_EBM.

Gupta, M. 2007. “Does Evidence-Based Medicine Apply to Psychiatry?” *Theoretical Medicine and Bioethics* 28 (2): 103–120. doi:10.1007/s11117-007-9029-x.

Guyatt, G. H., A. D. Oxman, G. E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H. J. Schünenmann. 2008. “GRADE: An Emerging Consensus on Rating Quality of Evidence and Strength of Recommendations.” *BMJ: British Medical Journal* 336 (7650): 924–926. doi:10.1136/bmj.39489.470347.AD.

Hempel, C. G. 1965. “Science and Human Values.” In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, edited by C. G. Hempel, 81–96. New York: Free Press.

Hill, A. B. 1951. “The Clinical Trial.” *British Medical Bulletin* 7 (4): 278–282. doi:10.1093/oxfordjournals.bmb.a073919.

Hill, A. B. 1965. “The Environment and Disease: Association or Causation?” *Proceedings of the Royal Society of Medicine* 58 (5): 295–300.

Holman, B. 2019. “Philosophers on Drugs.” *Synthese* 196 (11): 4363–4390. doi:10.1007/s11229-017-1642-2.
Holman, B., and T. Wilholt. 2022. “The New Demarcation Problem.” *Studies in History and Philosophy of Science* 91: 211–220. doi:10.1016/j.shpsa.2021.11.011.

Holtum, S., and V. Huet. 2014. “The MATISSE Trial—A Critique: Does Art Therapy Really Have Nothing to Offer People With a Diagnosis of Schizophrenia?” *SAGE Open* 4 (2): 215824014532930. doi:10.1177/215824014532930.

Howick, J. 2011a. “Exposing the Vanities – and a Qualified Defense – of Mechanistic Reasoning in Health Care Decision Making.” *Philosophy of Science* 78 (5): 926–940. doi:10.1086/662561.

Howick, J. 2011b. *The Philosophy of Evidence-Based Medicine*. Wiley-Blackwell.

Howick, J., P. Glasziou, and J. K. Aronson. 2009. “The Evolution of Evidence Hierarchies: What Can Bradford Hill’s “Guidelines for Causation” Contribute?” *Journal of the Royal Society of Medicine* 102 (5): 186–194. doi:10.1258/jrsm.2009.090020.

Howick, J., and T. Hoffmann. 2018. “How Placebo Characteristics Can Influence Estimates of Intervention Effects in Trials.” *CMAJ: Canadian Medical Association Journal = Journal de L’Association Medicale Canadienne* 190 (30): E908–E911. doi:10.1503/cmaj.171400.

Howick, J., R. K. Webster, J. L. Rees, R. Turner, H. Macdonald, A. Price, et al. 2020. “TIDieR-Placebo: A Guide and Checklist for Reporting Placebo and Sham Controls.” *PLoS Medicine* 17 (9): e1003294. doi:10.1371/journal.pmed.1003294.

Illari, P. M. 2011. “Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis.” *International Studies in the Philosophy of Science* 25 (2): 139–157. doi:10.1080/02698595.2011.574856.

Illari, P., and F. Russo. 2014. *Causality: Philosophical Theory Meets Scientific Practice*. Oxford University Press.

Jerkert, J. 2015. “Negative Mechanistic Reasoning in Medical Intervention Assessment.” *Theoretical Medicine and Bioethics* 36 (6): 425–437. doi:10.1007/s11107-015-9348-2.

John, S. 2015. “Inductive Risk and the Contexts of Communication.” *Synthese* 192 (1): 79–96. http://www.jstor.org/stable/247044882019.

John, S. 2019. “Science, Truth and Dictatorship: Wishful Thinking or Wishful Speaking?” *Studies in History and Philosophy of Science Part A* 78: 64–72. doi:10.1016/j.shpsa.2018.12.003.

Jukola, S. 2019. “On the Evidentiary Standards for Nutrition Advice.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 73: 1–9. doi:10.1016/j.shpsc.2018.05.007.

Kaptchuk, T. J. 2002. “Acupuncture: Theory, Efficacy, and Practice.” *Annals of Internal Medicine* 136 (5): 374–383. doi:10.7326/0003-4819-136-5-200203050-00010.

Kirsch, I., B. Wampold, and J. Kelley. 2015. “Controlling for the Placebo Effect in Psychotherapy: Noble Quest or Tilting at Windmills?” *Psychology of Consciousness: Theory, Research, and Practice* 3. doi:10.1037/cns0000065.

Kotsovilis, S., and L. H. Slim. 2012. “Periodontal and Cardiovascular Diseases: Statistical or Causal Association? A Review and Analysis Using Hill’s Criteria for Causation.” *Canadian Journal of Dental Hygiene* 46 (2): 131–142.

La Caze, A., and M. Colyvan. 2017. “A Challenge for Evidence-Based Policy.” *Axiomathes* 27 (1): 1–13. doi:10.1007/s10156-016-9291-5.

Landes, J., B. Osimani, and R. Poellinger. 2018. “Epistemology of Causal Inference in Pharmacology.” *European Journal for Philosophy of Science* 8 (1): 3–49. doi:10.1007/s13194-017-0169-1.

Lee, A., S. K. Chan, and L. T. Fan. 2015. “Stimulation of the Wrist Acupuncture Point PC6 for Preventing Postoperative Nausea and Vomiting.” *Cochrane Database of Systematic Reviews* 11: Article 11. doi:10.1002/14651858.CD003281.pub4.

Lee, A., and L. T. Fan. 2009. “Stimulation of the Wrist Acupuncture Point P6 for Preventing Postoperative Nausea and Vomiting.” *Cochrane Database of Systematic Reviews* 2: Article 2. doi:10.1002/14651858.CD003281.pub3.
Leuridan, B., and E. Weber. 2011. “The IARC and Mechanistic Evidence.” In Causality in the Sciences, edited by P. M. Illari, F. Russo, and J. Williamson, 91–109. Oxford University Press. http://hdl.handle.net/1854/LU-1223153.

Lilienfeld, S. O., L. A. Ritschel, S. J. Lynn, R. L. Cautin, and R. D. Latzman. 2014. “Why Ineffective Psychotherapies Appear to Work: A Taxonomy of Causes of Spurious Therapeutic Effectiveness.” Perspectives on Psychological Science 9 (4): 355–387. doi:10.1177/1745691614535216.

Lillie, E. O., B. Patay, J. Diamant, B. Issell, E. J. Topol, and N. J. Schork. 2011. “The N-of-1 Clinical Trial: The Ultimate Strategy for Individualizing Medicine?” Personalized Medicine 8 (2): 161–173. doi:10.2217/pme.11.7.

Lubowitz, J. H., M. T. Provencher, and G. G. Poehling. 2014. “Congratulations and Condemnations: Level I Evidence Prize for Femoral Tunnel Position in ACL Reconstruction, and AAOS Clinical Practice Guidelines Miss the Mark-Again.” Arthroscopy: The Journal of Arthroscopic & Related Surgery 30 (1): 2–5. doi:10.1016/j.arthro.2013.11.008.

Luján, J. L., and O. Todt. 2020. “Standards of Evidence and Causality in Regulatory Science: Risk and Benefit Assessment.” Studies in History and Philosophy of Science Part A 80: 82–89. doi:10.1016/j.shpsa.2019.05.005.

Machamer, P., L. Darden, and C. F. Craver. 2000. “Thinking About Mechanisms.” Philosophy of Science 67 (1): 1–25.

McAlister, F. A. 2008. “The “Number Needed to Treat” Turns 20 — and Continues to Be Used and Misused.” Canadian Medical Association Journal 179 (6): 549–553. doi:10.1503/cmaj.080484.

Naudat, F., A. S. Maria, and B. Falissard. 2011. “Antidepressant Response in Major Depressive Disorder: A Meta-Regression Comparison of Randomized Controlled Trials and Observational Studies.” PLoS One 6 (6): e20811. doi:10.1371/journal.pone.0020811.

NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines) With NCCN Evidence Blocks™. 2021. https://www.nccn.org/guidelines/guidelines-with-evidence-blocks.

Neal, J. P. 2019. “When Causal Specificity Does Not Matter (Much): Insights from HIV Treatment.” Philosophy of Science 86 (5): 836–846. doi:10.1086/705510.

Nowinski, C. J., D. M. Miller, and D. Cella. 2017. “Evolution of Patient-Reported Outcomes and Their Role in Multiple Sclerosis Clinical Trials.” Neurotherapeutics 14 (4): 934–944. doi:10.1007/s13311-017-0571-6.

Parkkinnen, V.-P., C. Wallmann, M. Wilde, B. Clarke, P. Illari, M. P. Kelly, C. Norell, F. Russo, B. Shaw, and J. Williamson. 2018. Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures. Springer International Publishing. doi:10.1007/978-3-319-94610-8

Peer Review & Public Comments and AAOS Responses to “American Academy of Orthopaedic Surgeons. Treatment of Osteoarthritis of the Knee – 2nd Edition Evidence-Based Clinical Practice Guideline. 2012. https://www.aaos.org/quality/quality-programs/lower-extremity-programs/osteoarthritis-of-the-knee/.

Perneger, T. V., and T. Agoritsas. 2011. “Doctors and Patients’ Susceptibility to Framing Bias: A Randomized Trial.” Journal of General Internal Medicine 26: 1411–1417. doi:10.1007/s11606-011-1810-x.

Peters, E., N. Dieckmann, A. Dixon, J. H. Hibbard, and C. K. Mertz. 2007. “Less Is More in Presenting Quality Information to Consumers.” Medical Care Research and Review 64 (2): 169–190. doi:10.1177/10775587070640020301.

Phillips, C. V., and K. J. Goodman. 2006. “Causal Criteria and Counterfactuals; Nothing More (or Less) Than Scientific Common Sense.” Emerging Themes in Epidemiology 3: 5. doi:10.1186/1742-7622-3-5.

Pickering, G. W. 1949. “The Place of the Experimental Method in Medicine.” Proceedings of the Royal Society of Medicine 42 (4): 229–234.

Plutynski, A. 2017. “Safe or Sorry? Cancer Screening and Inductive Risk.” In Exploring Inductive Risk: Case Studies of Values in Science, edited by K. C. Elliott and T. Richards. Oxford University Press. doi:10.1093/acprof:oso/9780190467715.001.0001
Post, P. N., H. de Beer, and G. H. Guyatt. 2013. “How to Generalize Efficacy Results of Randomized Trials: Recommendations Based on a Systematic Review of Possible Approaches.” *Journal of Evaluation in Clinical Practice* 19 (4): 638–643. doi:10.1111/j.1365-2753.2012.01888.x.

Rehfuess, E. A., and E. A. Akl. 2013. “Current Experience With Applying the GRADE Approach to Public Health Interventions: An Empirical Study.” *BMC Public Health* 13 (1): 1–13. doi:10.1186/1471-2458-13-9.

Reiss, J. 2012. “Causation in the Sciences: An Inferentialist Account.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (4): 769–777. doi:10.1016/j.shpsc.2012.05.005.

Reiss, J. 2015. “A Pragmatist Theory of Evidence.” *Philosophy of Science* 82 (3): 341–362. doi:10.1086/681643.

Resnik, D. B. 2017. “Dual-Use Research and Inductive Risk.” In *Exploring Inductive Risk: Case Studies of Values in Science*, edited by K. C. Elliott and T. Richards. Oxford University Press. doi:10.1093/acprof:oso/9780190467715.003.0004

Rothman, K. J., and S. Greenland. 2005. “Causation and Causal Inference in Epidemiology.” *American Journal of Public Health* 95 (S1): S144–S150. doi:10.2105/AJPH.2004.059204.

Rudner, R. 1953. “The Scientist qua Scientist Makes Value Judgments.” *Philosophy of Science* 20: 1–6. doi:10.1086/287231.

Russo, F., and J. Williamson. 2007. “Interpreting Causality in the Health Sciences.” *International Studies in the Philosophy of Science* 21 (2): 157–170. doi:10.1080/02698590701498084.

Russo, F., and J. Williamson. 2011a. “Epistemic Causality and Evidence-Based Medicine.” *History and Philosophy of the Life Sciences* 33 (4): 563–581.

Russo, F., and J. Williamson. 2011b. “Generic Versus Single-Case Causality: The Case of Autopsy.” *European Journal for Philosophy of Science* 1 (1): 47–69. doi:10.1007/s13194-010-0012-4.

Rudner, R.1953. “The Scientist qua Scientist Makes Value Judgments.” *Philosophy of Science* 20: 1–6. doi:10.1086/287231.

Russo, F., and J. Williamson. 2007. “Interpreting Causality in the Health Sciences.” *International Studies in the Philosophy of Science* 21 (2): 157–170. doi:10.1080/02698590701498084.

Russo, F., and J. Williamson. 2011a. “Epistemic Causality and Evidence-Based Medicine.” *History and Philosophy of the Life Sciences* 33 (4): 563–581.

Russo, F., and J. Williamson. 2011b. “Generic Versus Single-Case Causality: The Case of Autopsy.” *European Journal for Philosophy of Science* 1 (1): 47–69. doi:10.1007/s13194-010-0012-4.

Sackett, D. L., W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. 1996. “Evidence Based Medicine: What it Is and What it Isn’t.” *BMJ* 312: 71. doi:10.1136/bmj.312.7023.71.

Savulescu, J., K. Warłowska, and A. Carr. 2016. “Randomised Placebo-Controlled Trials of Surgery: Ethical Analysis and Guidelines.” *Journal of Medical Ethics* 42 (12): 776. doi:10.1136/medethics-2015-103333.

Schünemann, H., S. Hill, G. Guyatt, E. A. Akl, and F. Ahmed. 2011. “The GRADE Approach and Bradford Hill’s Criteria for Causation.” *Journal of Epidemiology and Community Health* 65 (5): 392–395. doi:10.1136/jech.2010.119933.

Shi, Y., T. Xu, Q. Chen, J. Wu, Y. Zhong, S. Song, Y. Chen, W. Gao, and L. Zhao. 2019. “Acupuncture for Radiotherapy-Induced Nausea and Vomiting: A Systematic Review Protocol.” *Medicine* 98 (24): e16027. doi:10.1097/MD.0000000000016027.

Shimonovich, M., A. Pearce, H. Thomson, K. Keyes, and S. V. Katikireddi. 2021. “Assessing Causality in Epidemiology: Revisiting Bradford Hill to Incorporate Developments in Causal Thinking.” *European Journal of Epidemiology* 36 (9): 873–887. doi:10.1007/s10654-020-00703-7.

Smith, A. D., H. Refsum, T. Bottiglieri, M. Fenech, B. Hooshmand, A. McCaddon, J. W. Miller, I. H. Rosenberg, and R. Obeid. 2018. “Homocysteine and Dementia: An International Consensus Statement.” *Journal of Alzheimer’s Disease: JAD* 62 (2): 561–570. doi:10.3233/JAD-171042.

Solomon, M. 2015. *Making Medical Knowledge*. Oxford University Press.

Stang, A., C. Poole, and R. Bender. 2010. “Common Problems Related to the Use of Number Needed to Treat.” *Journal of Clinical Epidemiology* 63 (8): 820–825. doi:10.1016/j.jclinepi.2009.08.006.

Steele, K. 2012. “The Scientist qua Policy Advisor Makes Value Judgments.” *Philosophy of Science* 79 (5): 893–904. doi:10.1086/66784.

Stegenga, J. 2011. “Is Meta-analysis the Platinum Standard of Evidence?” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 42 (4): 497–507. doi:10.1016/j.shpsc.2011.07.003.

Stegenga, J. 2014. “Down With the Hierarchies.” *Topoi* 33: 313–322. doi:10.1007/s11245-013-9189-4.
Stegenga, J. 2015. “Measuring Effectiveness.” Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 54: 62–71. doi:10.1016/j.shpsc.2015.06.003.

Stegenga, J. 2017. “Drug Regulation and the Inductive Risk Calculus.” In Exploring Inductive Risk: Case Studies of Values in Science, edited by K. C. Elliott and T. Richards. Oxford University Press. doi:10.1093/acprof:oso/9780190467715.003.0002

Stiles, W. B. 2009. “Responsiveness as an Obstacle for Psychotherapy Outcome Research: It’s Worse Than You Think.” Clinical Psychology: Science and Practice 16 (1): 86–91. doi:10.1111/j.1468-2850.2009.01148.x.

Stiles, W. B., M. Barkham, J. Connell, and J. Mellor-Clark. 2008. “Responsive Regulation of Treatment Duration in Routine Practice in United Kingdom Primary Care Settings: Replication in a Larger Sample.” Journal of Consulting and Clinical Psychology 76 (2): 298–305. doi:10.1037/0022-006X.76.2.298.

Tresker, S. Forthcoming. “Treatment Effectiveness, Generalizability, and the Explanatory/Pragmatic-Trial Distinction.” Synthese.

U.S. Department of Health, Education, and Welfare. 1964. Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service. PHS Publication No. 1103. Washington DC: U.S. Department of Health, Education, and Welfare, Public Health Service, Center for Disease Control.

Uttley, L., M. Stevenson, A. Scope, A. Rawdin, and A. Sutton. 2015. “The Clinical and Cost Effectiveness of Group Art Therapy for People with Non-Psychotic Mental Health Disorders: A Systematic Review and Cost-Effectiveness Analysis.” BMC Psychiatry 15: 151. doi:10.1186/s12888-015-0528-4.

Walker, M. J., and W. Rogers. 2014. “What Can Feminist Epistemology Do for Surgery?” Hypatia 29 (2): 404–421. doi:10.1111/hyap.12052.

Ward, A. C. 2009. “The Role of Causal Criteria in Causal Inferences: Bradford Hill’s ‘Aspects of Association.’” Epidemiologic Perspectives & Innovations: EP + I 6: 2. doi:10.1186/1742-5573-6-2.

Wilde, M., and V.-P. Parkkinen. 2019. “Extrapolation and the Russo–Williamson Thesis.” Synthese 196 (8): 3251–3262. doi:10.1007/s11229-017-1573-y.

Wilholt, T. 2009. “Bias and Values in Scientific Research.” Studies in History and Philosophy of Science Part A 40 (1): 92–101. doi:10.1016/j.shpsa.2008.12.005.

Williamson, J. 2019. “Establishing Causal Claims in Medicine.” International Studies in the Philosophy of Science 32 (1): 33–61. doi:10.1080/02698595.2019.1630927.

Williamson, J. 2021. “Establishing the Teratogenicity of Zika and Evaluating Causal Criteria.” Synthese 198: 2505–2518. doi:10.1007/s11229-018-1866-9.

Wood, C. 2013. “In the Wake of the Matisse RCT: What About Art Therapy and Psychosis?” International Journal of Art Therapy 18 (3): 88–97. doi:10.1080/17454832.2013.850104.

Woodward, J. 2010. “Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation.” Biology & Philosophy 25: 287–318.

Zinman, B., C. Wanner, J. M. Lachin, D. Fitchett, E. Bluemki, S. Hantel, M. Mattheus, et al. 2015. “Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes.” New England Journal of Medicine 373 (22): 2117–2128. doi:10.1056/NEJMoa1504720.