Studies in History and Philosophy of Science 103 (2024) 95–104

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

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

Evidence of mechanisms in evidence-based policy

Saúl Pérez-González

Department of Philosophy, University of Valencia, Av. Blasco Ibáñez 30, Valencia, 46010, Spain

ARTICLE INFO

Keywords:
Evidence
Mechanism
Policy
Pluralism
Extrapolation

ABSTRACT

Evidence-based policy has achieved great relevance in policy-making and social research. Nonetheless, over the past few years, several problematic aspects of this approach have been identified. This paper discusses whether, and to what extent, evidence of mechanisms could contribute to addressing certain difficulties faced by evidence-based policy. I argue that it could play a crucial role in the assessment of the efficacy of interventions, the extrapolation of interventions to target populations, and the identification of side effects. For analysing the potential contribution of evidence of mechanisms, the previous debate on the pluralist approach to evidence-based medicine is taken as reference.

1. Introduction

The evidence-based policy (EBP) movement emerged between the late 1990s and the early 2000s. The central idea of this movement is that policy-making should be guided by the best available evidence. In the development of EBP, the establishment of the Campbell Collaboration in 1999 was a crucial milestone (Littell & White, 2018). This institution aims to produce systematic reviews and other evidence syntheses for policy-making. Another important event was the election of a new labour government, which embraced the evidence-based approach, in 1997 in the UK (Parsons, 2002).

EBP is inspired by the evidence-based medicine (EBM) approach. In fact, EBP was initially introduced as the direct application of the EBM methods to policy-making. In this sense, for instance, the Campbell Collaboration was developed taking the successful Cochrane Collaboration, which focuses on health care, as reference (Davies & Boruch, 2001; Littell & White, 2018). The guidance of EBM is explicitly acknowledged in most EBP white papers.

Over the last two decades, EBP has become a popular approach in social research. It has been applied to diverse areas such as development economics, crime prevention, education, housing policy, and criminal justice. Furthermore, EBP has achieved considerable relevance and influence in the US, the EU, and the UK. In 2013, for example, the British government established the What Works Network. This network aims to expand and consolidate the evidence-based approach in diverse areas of social policy (see Cabinet Office, 2013). It integrates several centres such as The What Works Centre for Local Economic Growth, The Education Endowment Foundation, and The What Works Centre for Wellbeing. Nonetheless, in the last years, several philosophers and social scientists have argued that the current EBP approach faces important difficulties and that reforms are needed. Those difficulties are usually related to the limitations of the evidence provided by the methods highly valued by EBP.

Some authors have suggested that the pluralist approach to evidence, originally developed in the framework of EBM, should be extended to EBP (Clarke, Gillies, Illari, Russo, & Williamson, 2013, 2014; Shan & Williamson, 2021).[1] They consider that evidence of mechanisms could play a prominent role in policy-making. Nonetheless, those tentative proposals provide no systematic analysis of the potential contribution of evidence of mechanisms in EBP. Furthermore, they do not consider the particularities of policy-making and its problems. Previous debate on evidence (and its use) in EBP is not properly taken into account.

The aim of this paper is to analyse whether, and in which sense, evidence of mechanisms could contribute to addressing the difficulties faced by EBP. For this purpose, I will take into account policy literature on evidence-based practice. The structure of the paper is as follows. Section 2 presents the main aspects of EBP and introduces some relevant problems faced by it. Section 3 examines the pluralist approach to EBM,

E-mail address: saul.perez@uv.es.

[1] As pointed out by one of the reviewers, realist synthesis has also advocated a broader approach towards evidence and evidence-gathering methods in EBP (Pawson, Greenhalgh, Harvey, & Walshe, 2005; Wong et al., 2013). This model of research synthesis "seeks to unpack the mechanism of *how* complex programmes work (or *why* they fail) in particular contexts and settings" (Pawson et al., 2005, p. 21). It provides guidelines and procedures for collecting, assessing, and synthesising evidence. Nonetheless, there are relevant divergences between realist synthesis and evidential pluralism. They differ in aspects such as the scope and the role of theory.

https://doi.org/10.1016/j.shpsa.2023.11.006
Received 17 March 2023; Received in revised form 20 November 2023; Accepted 30 November 2023
Available online 14 December 2023

which has assessed and vindicated the role of evidence of mechanisms in the health sciences. The following three sections, on the basis of that debate, discuss how evidence of mechanisms could contribute to addressing certain relevant problems of EBP. They focus on the assessment of the efficacy of interventions (section 4), the extrapolation of interventions to target populations (section 5), and the identification of side effects (section 6). Afterwards, section 7 identifies and analyses some challenges encountered in introducing evidence of mechanisms into EBP. Finally, section 8 concludes.

2. Evidence-based policy

The core principle of EBP is that the development, implementation, and evaluation of policies and social interventions should be guided by the best available evidence (Baron, 2018; Haskins, 2018; Head, 2008). It is argued that decision-making should be anchored by contemporary social and behavioural sciences. In other words, it should promote those interventions or approaches that are known to work. Following the evidence is considered to be the best strategy to ensure that what is done is worthwhile and well-implemented. EBP aims to avoid biases, flawed reasoning, and misplaced goodwill in decision-making, which have been responsible for many undesired outcomes in the past.

The application of the general principle of EBP is made feasible by the introduction of practical guidelines. Those guidelines usually include hierarchies of evidence, which were initially introduced in the framework of EBM, that rank the diverse evidence-gathering methods. The weight given to a particular piece of evidence depends on the position in the hierarchy of the method through which it has been obtained. A representative example of a hierarchy of evidence is provided by the Scottish Intercollegiate Guidelines Network (SIGN) (see Table 1).

Randomised controlled trials (RCTs) and studies based on them (e.g., systematic reviews of RCTs) are placed on the upper part of hierarchies of evidence. In an RCT, a trial sample, which is usually a sample from the population of interest, is divided into two wings: intervention group(s) and control group(s). Members of the trial sample (e.g., schools) are randomly assigned to one of those groups. In intervention groups, an intervention under examination is implemented, while, in control groups, a standard procedure or a placebo is applied. RCTs are usually presented as the 'gold standard' (Baron, 2018; Haynes, Goldacre, & Torgerson, 2012). Evidence provided by RCTs and studies based on them is considered the most reliable for establishing the efficacy of an intervention. Randomisation is argued to be able to control for both known and unknown confounders. When individuals are randomly allocated, the control and the intervention groups are likely to resemble in all relevant respects (except for the intervention). This includes factors that are known to be relevant, but also factors whose influence is unsuspected. Consequently, observed differences in outcome could be attributed to the intervention.

In hierarchies of evidence, RCTs are followed by the other evidence-gathering methods according to their alleged reliability. Case-control and cohort studies are usually placed after RCTs. Non-analytic studies (e.g., case reports) are placed at lower levels of the hierarchies. Finally, expert opinion and formal consensus are considered very unreliable and are located at the bottom. It is argued that this kind of evidence may be influenced by personal values (e.g., political views) and cognitive biases (e.g., confirmation biases). Furthermore, expertise may be difficult to define and demarcate. Regarding the combination or integration of the evidence provided by those methods, hierarchies usually include no guidelines. When there is a disagreement among diverse kinds of evidence, the standard recommendation is simply to follow the higher-ranked available evidence.

Table 1
Hierarchy of evidence from SIGN (2011, p. 51).

Levels of evidence
1++ High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+ Well conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias
1 - Meta-analyses, systematic reviews, or RCTs with a high risk of bias
2++ High quality systematic reviews of case control or cohort studies
2+ Well conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
2 - Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal
3 Non-analytic studies, eg case reports, case series
4 Expert opinion

To illustrate the EBP approach, let me briefly introduce the debate about the cost-sharing of health goods. In public health, a relevant debate concerns the optimal subsidy level for health goods, particularly for those goods that, like bed nets, require active usage by the owner (James et al., 2006). It is often considered that cost-sharing—charging a subsidized positive price—is necessary to avoid wasting resources on those people that do not need the product or do not use it appropriately. Nonetheless, some authors have argued that cost-sharing has many negative consequences and that free distribution of health products is the most reasonable strategy (see, for instance, Gilson & McIntyre, 2005).

To explore the trade-offs between cost-sharing and free distribution of health goods, Cohen and Dupas (2010) conducted an RCT in prenatal clinics in western Kenya. In the trial, the price at which 20 clinics sold long-lasting insecticide-treated bed nets (ITNs) to prevent malaria infection in pregnant women was randomised. Four clinics served as a control group and continued their regular activity selling ITNs at 50 Kenyan shillings (Ksh). The other clinics were provided with ITNs and instructed to assign a price of 0 Ksh, 10 Ksh, 20 Ksh, or 40 Ksh. Cohen and Dupas' experiment found no evidence that cost-sharing reduces wastage: prenatal clients that paid a positive price for ITNs were not more likely to put ITNs to intended use. Nonetheless, the experiment found that cost-sharing significantly dampens demand and, consequently, reduces the program's coverage. They estimated that each 10 Ksh increase in the price of ITNs leads to a 20% decline in weakly sales. On the bases of those results, the authors argued that "cost-sharing is at best marginally more cost-effective than free distribution, but [because it achieves a higher coverage] free distribution leads to many more lives saved" (Cohen & Dupas, 2010, p. 41).

In addition to this influential work by Cohen and Dupas, other studies (including several RCTs) about cost-sharing have been conducted. Those studies have provided similar results about the effects of cost-sharing on the use of health goods (see, for example, Lagarde & Palmer, 2011). Influenced by this evidence, a new consensus has emerged among global health actors (e.g., intergovernmental organisations): it is now widely accepted that user fees are not an appropriate tool (Robert & Ridde, 2013).

Over the past few years, several problematic aspects of EBP have been identified by both philosophers and social scientists. Many of those problems are related to the limitations of the evidence provided by RCTs and other methods highly valued by hierarchies of evidence. Those methods face important difficulties in assessing the efficacy of relevant policies and interventions (Gamoran, 2018; Gelman, 2018; Kemm, 2006; Khosrowi & Reiss, 2019; La Caze & Colyvan, 2017; Montuschi, 2009). They are also hardly adequate to support the extrapolation (or generalisation) of causal claims from the study populations (Cartwright, 2012; Cartwright & Hardie, 2012; Deaton & Cartwright, 2018; Knox, Hill, & Berlin, 2018). Other difficulties are related to the combination of evidence provided by diverse sources (Cartwright, Goldfinch, & Howick, 2009) and the identification of side effects (La Caze & Colyvan, 2017).

There are also important concerns not directly related to causal inferences. Some of those concerns refer to the ethical dimension of EBP. Although EBP is considered to involve value neutrality, it actually favours certain issues and policies (Khosrowi, 2019). Furthermore, in EBP, field experiments rarely include informed consent and equipoise (Lisciandra, 2020). Some other concerns are related to the path from

scientific evidence to policy decision-making (Drèze, 2018; Gamoran, 2018). Research findings are just one of many sources of information and decision-makers often have low awareness of its details. Moreover, the research sector does not always pay enough attention to decision--makers' needs.

As a result of those criticisms, there is an ongoing debate among practitioners about the status of the current EBP approach (French, 2019). Some authors consider that those problems result from poor applications of the EBP principles and that the theoretical bases of the program should remain unchanged (e.g., Gluckman & Wilsdon, 2016). Nonetheless, most authors acknowledge that the current EBP has significant caveats that should not be neglected. Among these authors, positions range from those who call for minor adjustments (e.g., a broader approach to evidence) to those who propose to abandon the evidence-based approach.

In sum, EBP holds that policy-making should be guided by the best available evidence. Evidence is evaluated based on hierarchies of evidence that rank evidence-gathering methods according to their alleged reliability. In those hierarchies, the highest levels are occupied by RCTs and studies based on them. Nonetheless, in the last few years, it has been shown that evidence provided by the methods highly valued by EBP has important limitations. In the context of EBM, it has been argued that evidence of mechanisms can contribute to addressing similar difficulties. In the next section, I will analyse evidence of mechanisms and its contribution in EBM.

3. Evidence of mechanisms in EBM

In the framework of EBM, the approach that EBP mirrors, concerns have also been raised regarding the limitations of the evidence provided by RCTs and the other methods highly valued by hierarchies of evidence. It has been argued that this kind of evidence faces important difficulties in crucial tasks such as establishing causal relationships (Worrall, 2007) or extrapolating causal claims from animal models (Steel, 2008).

In EBM, in response to those concerns, some authors have stressed the relevance of evidence of mechanisms and argued that it should play a more prominent role in the assessment of causal claims (Anjum, Copeland, & Rocca, 2020; Aronson, Caze, Kelly, Parkkinen, & Williamson, 2018; Clarke, Gillies, Illari, Russo, & Williamson, 2014; Gillies, 2018; Parkkinen et al., 2018; Russo & Williamson, 2007). Those authors advocate a pluralist approach that places evidence of mechanisms alongside with other kinds of evidence such as evidence of correlations provided by experimental studies. In this framework, evidence of mechanisms is understood as evidence about the existence or the properties of mechanisms in the domain of inquiry (Illari, 2011; Williamson, 2019). Regarding the notion of mechanism, a broad view is usually adopted: a mechanism can be a complex system, a causal process, or some combination of both (Parkkinen et al., 2018; Williamson, 2019). Furthermore, evidence of mechanisms can be obtained by means of diverse evidence-gathering methods such as *in vitro* experiments, case reports, or simulations. Evidence of mechanisms is not characterised by a particular evidence-gathering method, but by the object of evidence (Illari, 2011). In this sense, a mechanistic study is any study that investigates mechanisms in the domain of inquiry.

The development of evidential pluralism in EBM is strongly related to the Evidence-Based Medicine Plus project (EBM+) (Clarke et al., 2013, 2014; Parkkinen et al., 2018). The theoretical core of EBM+ is the Russo-Williamson thesis (Russo & Williamson, 2007). According to this epistemological thesis, both evidence of correlations and evidence of mechanisms are normally needed to establish a causal claim in medicine. It is argued that evidence of correlations and evidence of mechanisms complement each other. Evidence of correlations addresses the major weaknesses of evidence of mechanisms, and vice versa. Evidence of correlations may successfully assess the overall effect of the mechanisms acting on the considered outcome, but it can hardly deal with shortcomings related to confounding factors and non-causal correlations. On the other hand, evidence of mechanisms is appropriate to cope with the latter difficulties, but it cannot assess the overall effect of the involved mechanisms.

After the introduction of the Russo-Williamson thesis, several proposals that stress the relevance of evidence of mechanisms for the health sciences have been developed. It has been argued, for instance, that evidence of mechanisms should play a central role in the extrapolation of causal claims (Clarke et al., 2014; Wilde & Parkkinen, 2019). To extrapolate a causal claim from a study population to a target population, it is considered crucial to establish that the mechanisms at work in both populations are (sufficiently) similar in the relevant aspects. It has also been argued that evidence of mechanisms is relevant for other areas such as clinical practice (Tonelli & Williamson, 2020), risk assessment (Gillies, 2017), and drug development and approval (Aronson et al., 2018).

A clear example of the contribution of evidence of mechanisms in EBM is provided by the International Agency for Research on Cancer (IARC). The mission of the IARC is to prepare and publish scientific reviews and evaluations of evidence on the carcinogenicity of particular agents (e.g., environmental exposures) (IARC, 2019). On the basis of the available evidence, it aims to establish to what extent those agents can be considered carcinogenic to humans. The evaluation of agents is informed by three kinds of evidence: evidence of cancer in humans, evidence of cancer in experimental animals, and evidence of mechanisms. In that context, evidence of mechanisms mainly refers to information about toxicokinetics (e.g., metabolization) and processes of carcinogenesis (e.g., induction of epigenetic alterations). After separately analysing every individual study and assessing each stream of evidence in conjunction, the three kinds of evidence are integrated into an overall evaluation of the carcinogenicity of the agent. Then, according to that assessment, the agent is classified as carcinogenic to humans, probably carcinogenic to humans, possibly carcinogenic to humans, or non-classifiable as to its carcinogenicity to humans.

In the overall evaluation of agents, the main role of evidence of mechanisms is to guide the extrapolation of causal claims from animal models to humans (Leuridan & Weber, 2011). When evidence of cancer in humans is limited or inadequate, evidence of mechanisms is demanded to determine whether the agent is carcinogenic to humans. That role of evidence of mechanisms was crucial in the evaluation of benzo[*a*]pyrene (IARC, 2012), which has been discussed in detail by Wilde and Parkkinen (2019). Although no evidence of cancer in humans was available, benzo[*a*]pyrene was classified as carcinogenic to humans. The similarity between the mechanisms operating in experimental animals and those operating in humans enabled the extrapolation of the causal claim. It was considered that "[t]he strong and extensive experimental evidence for the carcinogenicity of benzo[*a*]pyrene in many animal species, supported by the consistent and coherent mechanistic evidence from experimental and human studies provide biological plausibility to support the overall classification of benzo[*a*]pyrene as a human carcinogen" (IARC, 2012, p. 138).

Summing up, in EBM (as in EBP), concerns have been raised regarding the limitations of the methods highly valued by the hierarchies of evidence. In order to address those concerns, some authors hold that evidence of mechanisms should play a more prominent role. It is argued that evidence of mechanisms is crucial for aims such as establishing and extrapolating causal claims. Following this pluralist approach to EBM, it has been argued that evidence of mechanisms could also play a relevant role in other fields such as econometrics (Moneta & Russo, 2014) and behavioural sciences (Grüne-Yanoff, 2016). In this sense, some authors have recently suggested that evidence of mechanisms could be a crucial resource for EBP (see, for example, Shan & Williamson, 2021). Nonetheless, those tentative proposals do not address the issue in detail. In the following sections, I will discuss whether, and in which sense, evidence of mechanisms could contribute to addressing certain relevant problems in EBP.

4. Efficacy of policy interventions

In EBP, RCTs are presented as a crucial method for establishing causal relationships and assessing the efficacy of policy interventions. They are not only placed at the highest levels of the hierarchies of evidence, but also provide the bases for other well-valued evidence-gathering methods such as meta-analyses and systematic reviews. The methodology of RCTs is considered to allow us to control for known and unknown confounding factors, minimise selection biases, and rule out chance. Nevertheless, it is often the case that evaluations of policy interventions cannot—or, at least, should not—rely on RCTs.

There are several reasons why RCTs may not be available or suitable for assessing a policy intervention. First, it is often the case that the relevant questions for policymakers cannot be adequately addressed by means of RCTs. Some important interventions, for instance, cannot be tested experimentally (e.g., changes in macroeconomic policies) (Khosrowi & Reiss, 2019). RCTs are also poorly suited when we are not just interested in the average effect of the intervention, but also in the distribution of the effect (Deaton, 2010; Gamoran, 2018; Khosrowi, 2019). Furthermore, RCTs are of limited help when there are many (different kinds of) variables of interest (La Caze & Colyvan, 2017). They can only compare a limited set of variables. Each variable of interest might be influenced by many factors, which should be balanced among populations. When the number of considered variables (and, consequently, relevant factors) increases, randomisation is less likely to result in balanced populations. This is particularly concerning when study populations are small. Second, in EBP, unlike in EBM, the causal knowledge required to design, conduct, and analyse RCTs is not always available (La Caze & Colyvan, 2017). Basic biological sciences such as biochemistry and physiology provide solid grounds for RCTs in EBM. However, in EBP, given the scarcity of well-established causal models, information about the processes underlying the trial is often sketchier. This may result in difficulties in designing placebo interventions, establishing appropriate markers, interpreting obtained results, etc. Third, it is very difficult to design and conduct good RCTs when the units of intervention are communities (e.g., neighbourhoods), rather than individuals (Kemm, 2006). Because of their complexity and small number, it is very unlikely that randomisation results in groups similar in all relevant aspects. And fourth, it may be unethical to test certain interventions by means of RCTs (Barrett & Carter, 2010). The randomised allocation of (scarce) resources, instead of its allocation according to people's needs, may give rise to ethical concerns.

In those cases in which RCTs are not available, EBP relies on non-randomised experimental studies or observational studies for assessing interventions. Those studies, unlike ideal RCTs, are susceptible to problems related to confounding factors and non-causal correlations. When a correlation between an intervention and a putative effect is observed, several interpretations are possible. Certainly, the considered intervention may cause and be responsible for the effect. Nonetheless, it may also be the case that some confounding factors are responsible for the correlation or that the correlation is just the result of chance (for an exhaustive list of explanations, see Williamson, 2019). Non-randomised experimental studies or observational studies can certainly establish the correlation between the intervention and the putative effect. However, they are unable to discard other possible interpretations and establish a causal relationship between them.

I consider that, in these scenarios, evidence of mechanisms could play a crucial role and help to discern whether the identified correlation is the result of a causal relationship between the intervention and the putative effect. Evidence of the mechanism (or mechanisms) through which the intervention produces the effect supports the causal interpretation of the correlation (Clarke et al., 2014; Russo & Williamson, 2007; Steel, 2004). This evidence provides details about the (organised) entities and activities that, as a result of the intervention, bring about the considered effect. For instance, it may be found that certain communication campaigns and increased use of public transport are linked by a mechanism of self-fulfilling prophecy (Biggs, 2009). Other examples of social mechanisms that may underlie an intervention are person-to-person transmission, vacancy chains, and coordinated action. Evidence of mechanisms also undermines alternative interpretations of the correlation, e.g., non-causal correlation resulting from chance. It should be noted that evidence of social mechanisms can be provided by diverse methods such as process tracing and case studies (Beach & Pedersen, 2019), which are often available in those scenarios in which RCTs cannot be conducted.

Regarding the assessment of policy interventions, evidence of mechanisms may also play a negative role and cast doubt on the efficacy of an intervention. The absence of evidence of the mechanisms through which the intervention brings about the putative effect undermines the causal interpretation of the identified correlation. To illustrate the negative role, consider the debate on 'climate wars' (Gilmore, Herzer Risi, Tennant, & Buhaug, 2018; Sakaguchi, Varughese, & Auld, 2017). Many studies have found a positive correlation between climate change (e.g., precipitation patterns) and violent conflicts (e.g., interstate wars). Furthermore, diverse intermediary mechanisms have been hypothesised (e.g., economic contraction). Nonetheless, evidence about the causal pathway through which climate change influences violent conflict is limited and inconclusive. This absence of evidence of mechanisms undermines causal interpretations of the relation and hinders the development of violence reduction policies.

When RCTs are available and suitable—i.e., when interventions can be tested using RCTs—difficulties related to confounders and non-causal correlations are less threatening. RCTs are designed to control for both known and unknown confounding factors and avoid chance. As Cartwright (2007) has argued, ideal RCTs are clinchers. When all the assumptions are met, they allow us to deduce causal claims about the study population. However, in EBP, real RCTs are usually far from the ideal. Relevant features such as (double) blind and use of a placebo are often absent. Consequently, in real RCTs, when a correlation is identified between an intervention and a putative effect, it may still be the case that the correlation is spurious. The considered intervention may not be responsible for the detected effect. The identified correlation may be the result of relevant differences among the causal factors in each population, systematic errors in the trial (e.g., confirmation biases), or even non-systematic errors (e.g., measurement error). Evidence of mechanisms, as it has been argued, can help to address those difficulties. It allows us to discern whether the identified correlation is the result of a causal relationship between the intervention and the putative effect.

In those cases in which policy interventions can be assessed by RCTs, evidence of mechanisms is also important for designing and assessing the trials. In EBP, it is often the case that units of analysis are complex, study populations are small, and randomisation is conducted only once. Therefore, important differences between the intervention and the control group may exist. Considering the mechanisms at work in both groups, it is possible to assess whether they resemble respect to the relevant causal factors. In this sense, evidence of mechanisms contributes to identifying (and avoiding) inaccurate RCTs that give rise to spurious correlations resulting from previous differences between populations. Furthermore, evidence of the mechanism of action of the intervention is crucial to identify what are those relevant causal factors regarding which the intervention and the control populations should resemble. This evidence illustrates how the intervention works and which factors are involved in that process. Another relevant task to which evidence of mechanisms contributes is establishing adequate markers for the effects of interest.

Evidence of mechanisms is also crucial for conducting good meta-analyses and systematic reviews. Meta-analyses and systematic reviews, even when they are based on RCTs, may be inaccurate and not provide reliable evidence about the efficacy of interventions. Meta-analyses may overlook important differences between the studies whose results are combined, e.g., differences regarding interventions (Kemm, 2006). Similarly, systematic reviews may not pay enough

attention to the circumstances in which the diverse studies have been conducted, and become a fallacious enumeration of very similar studies (Cowen, Virk, Mascarenhas-Keyes, & Cartwright, 2017). Evidence of mechanisms is helpful to address those problems. Evidence of the mechanisms through which the intervention works in the studied populations underlines the relevant causal factors that should be considered for addressing, comparing, and integrating the diverse studies.

In a nutshell, in EBP, evidence of mechanisms could contribute to establishing causal relationships and assessing the efficacy of policy interventions. When a correlation is identified between an intervention and a putative effect, a causal relationship between them cannot be unequivocally concluded. Many other interpretations of the correlation are in fact possible (e.g., non-causal correlation). Evidence of mechanisms can significantly contribute to discerning whether the identified correlation is the result of a causal relationship between the considered intervention and the effect. Evidence of mechanisms is also important for accurately conducting and assessing RCTs, meta-analyses, and systematic reviews.

5. Extrapolation of policy interventions

Policy-makers are usually interested in knowing whether certain policy interventions would work in the populations they are in charge of. Nonetheless, in EBP, the study population in which an intervention is tested is very rarely the target population of interest. Because of diverse reasons (ethical, economic, political, etc.), access to the target population is often limited. In most cases, the study population is just a sample of the target population or even a different population. This means that, in EBP, a crucial task is to extrapolate causal claims about the efficacy of interventions from study populations in which they are tested to target populations of interest. Nonetheless, simple extrapolation based on simple induction is not reliable (Gelman, 2018; Kemm, 2006). Even when the intervention has been shown to be effective in the study population, its efficacy in the target population cannot be taken for granted. An intervention may work in one population or context but not in a (slightly) different one.

RCTs and other methods advocated by the EBP framework face important difficulties in assessing the extrapolations of causal claims (Cartwright, 2009, 2012; Cartwright & Hardie, 2012; Deaton & Cartwright, 2018). Those methods may certainly provide relevant evidence of the efficacy of the intervention in the study population. However, evidence provided by them is hardly enough for supporting its efficacy in other scenarios. Hierarchies of evidence focus on internal validity and assess evidence-gathering methods on the basis of their reliability regarding the study population. In this sense, RCTs are highly valued because they are considered to maximise the degree of internal validity. Nonetheless, internal validity does not ensure external validity. Assuming a universal response and, consequently, understanding external validity as a consequence of internal validity is not reasonable when long and complex pathways are involved (Victora, Habicht, & Bryce, 2004). As Cartwright has claimed, there is "a long and tortuous road from learning that a policy works somewhere, which is the kind of claim an RCT can clinch, to correctly predicting that it will—or won't—work for you" (Cartwright, 2012, p. 988).

A well-known episode that exposes the difficulties faced by RCT-based extrapolations is the failure of the California Class Size Reduction program. In the 1980s, there was no consensus among lawmakers about the convenience of outlawing large classes. Past studies on class size showed mixed results and there were doubts about its cost-effectiveness. To study the effects of class-size reduction, the State of Tennessee funded Project STAR (for Student-Teacher Achievement Ratio). In the first phase of the project, an RCT to test the efficacy of class-size reduction was conducted (Mosteller, 1995). In 79 schools, kindergarten students and teachers were randomly assigned to classes large (22–25 students) or small (13–17 students). This first phase of the project had a duration of four years. During that period, students' math and reading skills were tested annually. The trial showed that reducing class size could significantly increase student learning in both maths and reading. Further studies also showed that low-income and minority students benefitted more from class-size reduction and that students assigned to smaller classes were more likely to attend college. In 1996, motivated in part by Project STAR, the State of California passed a law setting a maximum class size in kindergartens and providing $1 billion per annum to assist districts in class size reduction. Nonetheless, the intervention had little effect. In fact, no conclusive link was detected in California between class-size reduction and student achievement (Bohrnstedt & Stecher, 2002).

To extrapolate the efficacy of an intervention from a study population to a target population of interest, it is crucial to establish that they are similar enough regarding the relevant causal factors. Or, alternatively, that there is no particularity of the target population, which differentiates it from the study population, which severely undermines the efficacy of the intervention. I consider that evidence of mechanisms could play a relevant role in the assessment of the similarity between the study and the target population and, consequently, help to evaluate the extrapolation of causal claims about policy interventions. On the basis of evidence of mechanisms, we can examine the relevant mechanisms in the study population and whether, and to what extent, the mechanisms at work in the target population are similar to them (Steel, 2008). Regarding extrapolation, as regarding efficacy assessment (see section 4), evidence of mechanisms can play both a positive and a negative role. On the one hand, if the relevant mechanisms at work in the study and the target population are highly similar in the relevant aspects, the extrapolation of the causal claim is supported. On the other hand, if the relevant mechanisms at work in the study and the target population differ in relevant aspects, the extrapolation of the causal claim is undermined. Some social mechanisms that may differ from one population to another are education systems, cultural markets, or processes for reintegrating former convicts.

It should be noted that the mechanism-based assessment of the similarity between the study and the target population is not limited to the mechanism through which the intervention brings about the considered effect. It should also consider the other relevant mechanisms that may significantly influence the effect of interest (Williamson, 2019). Otherwise, there may be disturbing mechanisms that affect the outcome and, eventually, mask the influence of the considered mechanism of action in the target population.

In order to illustrate the potential contribution of evidence of mechanisms to the extrapolation of policy interventions, let me briefly discuss the difficulties related to necessary support factors (Cartwright, 2012; Cartwright & Hardie, 2012; Deaton & Cartwright, 2018). Policy intervention is rarely sufficient to produce the effect of interest. It usually requires the presence of certain support factors that allow it to operate properly (e.g., social structures). Consequently, even if the considered intervention has the capacity to operate in the target population, the absence of relevant support factors may prevent it to produce the desired effect (or even make it contribute to an undesired effect).

The failure of the California Class Size Reduction program, in fact, can be understood in terms of the absence of required support factors (Cartwright & Hardie, 2012, p. 65). To improve student achievement, class-size reduction requires certain support factors such as study space and qualified teachers. In absence of those factors, the intervention is unable to contribute to the targeted effect. Those required factors were present when the intervention was tested in Tennessee. Nonetheless, when the class-size reduction policy was introduced in California, several relevant support factors were absent. For example, given the large scale of the implementation, there was a shortage of qualified teachers to staff the smaller classes (and thousands of new and unskilled teachers were hired). The absence of required support factors undermined the contribution of class-size reduction and prevented it from improving students' performance.

Evidence of mechanisms can help to assess the presence of required

support factors and avoid unsuccessful extrapolations. First, evidence of the mechanism through which the intervention works in the study population provides crucial information about which are the relevant support factors that allow the intervention to operate. And second, considering the evidence of the mechanisms at work in the target population, it can be determined if the identified relevant support factors are present in that population.

In summary, evidence of mechanisms could contribute to assessing the extrapolation of causal claims about the efficacy of policy interventions from the study populations to the target populations of interest. Even when an intervention has been effective in a certain study population, it may not work in the target population. However, RCTs and other methods highly valued by hierarchies have important limitations in evaluating the extrapolation of causal claims. Evidence of mechanisms can identify the relevant causal factors in the study population and, based on this analysis, evaluate whether the study and the target populations are similar enough to justify the extrapolation of claims about the intervention.

6. Side effects of policy interventions

Over the last two decades, many concerns have been raised about the limitations of RCTs and other methods highly valued by hierarchies of evidence for identifying and anticipating undesirable side effects (Gillies, 2017; Rocca, Anjum, & Mumford, 2020; Vandenbroucke, 2006). In this sense, it has been claimed that, regarding side effects, evidence from observational studies is, at least, as good as evidence from RCTs (Osimani, 2013; Vandenbroucke, 2008). Regarding EBP, several authors have also stressed the relevance of these problems and the need of addressing them (Cartwright & Hardie, 2012; La Caze & Colyvan, 2017).

In EBP, in relation to undesirable side effects, RCTs have two main limitations. They face difficulties in (i) identifying side effects in the study population and (ii) anticipating side effects that, although do not affect the study population, would arise in the target population.

When a policy intervention is evaluated, RCTs have important difficulties for identifying the undesired side effects that affect the studied population. As it has been noted, RCTs are just able to control and compare a small set of variables of interest (La Caze & Colyvan, 2017). They are not suitable for simultaneously considering many diverse variables. Nonetheless, undesired effects may involve a wide variety of variables. In those cases in which side effects affect variables beyond the set under consideration, RCTs are likely to overlook them. Suppose, for instance, that an RCT is conducted to evaluate the flexible workplace practice of hot desking and its effect on employees' productivity. This kind of trial would probably be unable to identify some undesired effects related to employees' welfare (e.g., higher incidence of anxiety-related problems).

Furthermore, even when undesirable side effects that affect the study population involve some of the considered variables, RCTs may still face difficulties for identifying them. Real RCTs usually have important practical limitations (e.g., time limitations). Given those limitations, relevant side effects may not take place (or only take place partially) during the course of the trial. In those cases in which side effects do not arise during the trial, RCTs can hardly provide the basis for identifying them. Consider, for example, time limitations (Gillies, 2017). RCTs, which always have a specific endpoint, only monitor the considered variables until a certain point. At that point, the variables are compared, and the results of the trial are evaluated. Long-term side effects that take place after their endpoint cannot be identified by RCTs. Another practical limitation that may hinder the identification of undesirable effects is the limited number of participants. Some relevant side effects may be very rare and do not affect any member of the intervention group monitored in a particular RCT.

In EBP, RCTs also have difficulties for anticipating undesirable side effects that, although do not arise in the study population, would affect the target population. As it has been argued, the relevant causal factors in the study and the target population may differ in important aspects. When those differences are present, an intervention may produce certain undesirable effects in the target population but not in the study population. Nonetheless, RCTs do not provide the basis for anticipating those side effects before the introduction of the intervention in the target population.

To illustrate these limitations of RCTs, let me consider again the California Class Size Reduction program. In California, class size reduction produced undesired effects that did not take place in the trial conducted in Tennessee. A particularly relevant one was to hinder several activities that, independently of class size, contribute to students' performance (Cartwright & Hardie, 2012). In California, unlike in Tennessee, many schools did not have enough spare space for allocating the additional classes. In many cases, the required additional space was taken away to other activities such as music, athletics, or special needs. The resulting time reduction or absence of adequate space undermined the contribution made by those activities and had a negative effect on students' performance.

In the identification of side effects in the study population and the anticipation of side effects in target populations, as I will argue, evidence of mechanisms can also play an important role. First, evidence of the mechanism through which the intervention works in the study population can provide relevant information about side effects. This evidence can reveal undesirable effects that have taken place at some stage of the causal process. Paying attention to the mechanism of action is crucial to consider the diverse stages of the process and not overlook important effects. Moreover, evidence of the mechanism of action may provide clues for identifying potential side effects that, because of the practical limitations of the study, have not taken place. Even if an undesirable effect has not (completely) taken place, the presence of some supporting factors may make it expected. It may also be the case that the first stages of a causal process that is likely to bring about a side effect are observed.

And second, evidence of the mechanisms at work in the target population can significantly contribute to anticipating (new) undesirable side effects in that population. Taking into account the mechanism through which the intervention works, we can foresee side effects that would result from introducing the intervention in the mechanism complex of the target population. The mechanism of action may introduce causal factors that, in conjunction with factors already present in the target population, result in the production of undesirable effects. In an extremely polarised society, for example, an intervention that gives a decision-making role to a subgroup in which a certain faction is overrepresented may result in discriminatory behaviours. It can also be the case that, as in the California Class Size Reduction program, the mechanism through which the intervention works undermines another mechanism, which produces a desired effect, already present in the target population. In those scenarios, a detailed analysis of the mechanisms at work in the target population is likely to help us to anticipate the potential undesirable effects.

In sum, in EBP, evidence of mechanisms could contribute to identifying and anticipating undesirable side effects of policy interventions. When an intervention is tested, RCTs have important limitations for identifying side effects in the study population. They also face difficulties to anticipate side effects that, although did not affect the study population, would arise in target populations of interest. Evidence of mechanisms is crucial for addressing those limitations. Evidence of the mechanism of action of the intervention can provide information about the occurrence of undesirable side effects in the study population. Moreover, considering the mechanisms at work in the target population, it is also possible to anticipate side effects that would take place in that population.

7. Evidence of mechanisms: challenges for EBP

Evidence of mechanisms could contribute to address important limitations of current EBP. As it has been argued in the previous sections,

it is helpful to assess the efficacy of interventions, extrapolate causal claims, and identify side effects. Nonetheless, the introduction of evidence of mechanisms into EBP also faces important challenges. These challenges result from general limitations of evidence of mechanisms, but also from relevant divergences between EBM and EBP. In what follows, I will separately consider both types of challenges.

Evidence of mechanisms is not problem-free and also faces difficulties. One of the main problems of evidence of mechanisms is related to the limited information about relevant mechanisms. When a mechanism is identified in a population, the knowledge about it is often fragmentary. It is usually the case that only some components of the relevant mechanisms are known. Consequently, it may be complicated to identify the exact role played by a mechanism or anticipate how it would be affected by a certain change in a relevant causal factor. Another important difficulty faced by evidence of mechanisms is masking (Clarke et al., 2014). Even if some mechanisms are identified and scrutinised in a certain population, unknown disturbing mechanisms may also be present. Disturbing mechanisms would affect the outcome and, eventually, mask the influence of the considered mechanisms. In fact, those mechanisms could even interfere with the identified mechanisms and undermine their own contribution to the outcome. Other problems of evidence of mechanisms are related to mechanisms' irregular behaviour (Howick, Glasziou, & Aronson, 2010), absence of methods for comparing mechanisms (Reiss, 2010), and limitations for addressing quantitative questions (van Eersel, Koppenol-Gonzalez, & Reiss, 2019).

The problems faced by evidence of mechanisms, nevertheless, are not always as threatening. The relevance of those problems depends to a large extent on the role that evidence of mechanisms is supposed to play (Pérez-González & Iranzo, 2021). Consider, for example, the problem of masking. Masking is particularly ominous in the extrapolation of causal claims to target populations of interest (Howick, Glasziou, & Aronson, 2013; van Eersel et al., 2019). The access to target populations of interest, as has been noted, is often very limited. In some cases, only some of the mechanisms at work in the target populations are known. Consequently, even if several relevant mechanisms are identified in the study population and the presence of similar mechanisms is established in the target population, there may be unknown disturbing mechanisms at work in the target population. Those disturbing mechanisms would modify the outcome of the intervention and, as a result, the extrapolation of the causal claim would be erroneous.

Nonetheless, masking is less problematic when an undesired side effect is identified in the study population and an intervention is judged as hazardous. By analysing the mechanism through which an intervention produces the targeted effect, it is possible to identify the raise of undesired effects at some stage of the process. Certainly, as in extrapolation, it may be the case that there are unknown disturbing mechanisms in the population. Nevertheless, it is very unlikely that, because of the presence of those mechanisms, the judgement about the intervention is erroneous. Unknown disturbing mechanisms, to prevent the undesired side effect and undermine the judgement, should exactly compensate for the identified hazardous mechanism. However, given the complexity of most social mechanisms, that exact counterbalance is highly unlikely (Wagenaar, 2007). In most cases, disturbing mechanisms may just modify the identified undesired effect, e.g., making it more intense.

As advocates of evidential pluralism have argued, some problems of evidence of mechanisms could be addressed from a pluralistic framework (Clarke et al., 2014). Evidence of correlations can complement evidence of mechanisms and mitigate the abovementioned problems. Evidence of correlations, for instance, may provide the basis for addressing the problem of masking. It can measure the overall causal effect (i.e., the net outcome) of a certain intervention in a particular population. The measurement of the overall effect, which may deviate from the expected outcome given the known mechanisms, reveals whether there are disturbing mechanisms that modify the causal relation. Nonetheless, in EBP, the availability of the required complementary evidence of correlations cannot be taken for granted. In the relevant contexts, it is not always the case that (reliable) evidence of correlations is available. In particular, given the limited access to them, it may be difficult to obtain reliable evidence of correlations about target populations of interest. The limitations of evidence of mechanisms, therefore, are relevant in EBP and should be taken into account for the introduction of evidential pluralism.

I will now discuss some additional challenges resulting from the particularities of EBP. Evidential pluralists, when considering EBP, usually assume that a straightforward extension of the EBM's pluralism can be conducted. In this sense, Shan and Williamson say that "[m]ost obviously, the move from EBM to EBM + warrants an analogous move from present-day evidence-based policy (EBP) to EBP+" (2021, p. 7). Nonetheless, there are relevant divergences between EBM and EBP regarding both mechanisms and mechanistic studies. To integrate evidence of mechanisms in EBP, it is crucial to assess those divergences and take into account the particularities of EBP.

In the biomedical sciences, evidence of mechanisms usually refers to evidence of biological mechanisms. However, in EBP, most relevant mechanisms are social. While some general traits are shared by mechanisms across the sciences, there are relevant divergences between biological and social mechanisms. Biological mechanisms usually have a high degree of homogeneity. Most biological mechanisms are similar across individuals and populations. For instance, consider mechanisms of neurotransmitter release. The component entities (e.g., calcium ions) and activities (e.g., voltage rise) remain similar in most individuals. Social mechanisms, however, have a lower degree of homogeneity. There are usually relevant differences among social mechanisms across populations. Consider, for example, monetary transmission mechanisms. In many countries, there are mechanisms through which central banks (or similar institutions) influence employment, prices, and aggregate output. Nonetheless, there are important differences among monetary transmission mechanisms across countries. The mandates and dispositions of central banks, for instance, are diverse.

Another important divergence concerns organisation. Internal organisation is crucial in biological mechanisms. This organisation mainly involves spatial, temporal, and active aspects (Craver & Darden, 2013). Spatial organisation refers to aspects such as locations, sizes, or shapes; while temporal organisation includes orders, rates, and durations of stages. Active organisation denotes that component entities interact and make a difference in each other. Although internal organisation is also crucial in social mechanisms, it cannot be reduced to spatial, temporal, and active aspects. There are other relevant organisational aspects present in social mechanisms. In particular, norms usually play an important role. Social life is organised through norms that help to coordinate the activities of individuals and reduce conflict among them (Demeulenaere, 2021). Those norms can be formal, such as legal norms, or informal, such as social norms. In an agricultural market, for instance, social norms about (indirect) competition constrain the interactions among sellers.

Regarding mechanistic studies, there are also relevant divergences between EBM and EBP. Both fields differ concerning which evidence-gathering methods are usually available and employed to obtain evidence of mechanisms. In EBM, mechanistic studies can consist, among others, of "*in vitro* experiments, biomedical imaging, autopsy, established theory, animal experiments, and simulations" (Parkkinen et al., 2018, p. 14). Among them, methods that provide direct evidence of mechanisms are usually preferred. They allow us to directly observe (e.g., autopsy) or manipulate (e.g., *in vitro* experiments) mechanisms. To illustrate that point, let me refer again to the evaluation of benzo[*a*]pyrene by IARC (2012). To assess the mechanisms underlying benzo[*a*]pyrene-induced carcinogenesis, *in vitro* studies in cell lines and *in vivo* studies in animal models were reviewed. In this sense, evidence about the diolepoxide mechanism for benzo[*a*]pyrene was provided by studies in human and bovine bronchial explants. Nonetheless, methods employed in EBM to direct observation or manipulation of mechanisms

are rarely available in EBP (Wong, Westhorp, Pawson, & Greenhalgh, 2013). Consider, for example, *in vitro* experiments. Social mechanisms cannot be isolated from any social context and manipulated to study their components and properties.

In EBP, mechanistic studies often consist of agent-based models (ABMs) (Hedström & Ylikoski, 2010; Macy & Flache, 2009). Given the limitations for studying social mechanisms, social scientists have vindicated ABMs as a crucial tool. In this sense, regarding the study of mechanisms, Hedström and Ylikoski claim: "[g]iven the limitations of experimental methods and the complexity of social phenomena, computer simulations are important for this kind of endeavor" (Hedström & Ylikoski, 2010, p. 62). ABMs provide indirect evidence of social mechanisms. Relying on ABMs, we can test and assess hypotheses about the mechanisms underlying social macro-phenomena. When an ABM is able to generate the phenomenon of interest, the mechanistic hypothesis that informed it is reinforced.[2] Given their bottom-up nature, ABMs can be calibrated according to specific considerations about individuals, their properties, actions, and interactions. As an illustration, consider unintended residential segregation (Flache & de Matos Fernandez, 2021). To identify and study the mechanism responsible for this social phenomenon, ABMs (e.g., Schelling's model) have been crucial. They have shown that the mechanism of relocation cascade can generate segregated neighbourhoods.

Furthermore, in policy settings, compared to biomedical practice, more (and more diverse) perspectives are usually relevant for the development and assessment of interventions (see, for instance, Lohse & Canali, 2021). Those perspectives may involve, among others, economic, social, ecological, health, and legal aspects. For example, when lockdown policies were discussed during the COVID-19 pandemic, both economic and health aspects were considered. This plurality of perspectives raises a challenge for the study of mechanisms. Diverse perspectives may differ regarding their assessment of mechanisms. For instance, a mechanism that has no side effect from an economic perspective may have side effects from an ecological perspective. Likewise, the mechanisms at work in two populations may resemble from an economic perspective but differ from an ecological perspective.

8. Conclusion

According to EBP, policy-making should be guided by the best available evidence. Following EBM, evidence is ranked on the basis of hierarchies that place RCTs and studies based on them at the top. Since its emergence, EBP has gained great relevance in social research and policy-making. Nonetheless, over the past few years, several problematic aspects of EBP have been identified. Those problems are usually related to the limitations of RCTs and other methods highly valued by hierarchies of evidence. In this paper, I have argued that evidence of mechanisms could contribute to addressing some relevant limitations of the EBP framework. First, evidence of mechanisms helps to assess the efficacy of interventions. It allows us to discern whether a correlation between an intervention and certain output is the result of a causal relationship. Second, evidence of mechanisms supports the evaluation of extrapolations of policy interventions. It can assess the similarity between the study and the target population regarding the relevant causal factors. And third, evidence of mechanisms provides the bases for identifying and anticipating undesirable side effects of interventions.

Nonetheless, it is important to be cautious about evidence of mechanisms and its role in EBP. This evidence should not be considered a new and problem-free gold standard. Evidence of mechanisms, as any other kind of evidence, cannot be taken into account without prior analysis and evaluation. The quality of evidence of mechanisms may vary significantly. For example, the information about certain mechanism can range from very complete to fragmentary. Furthermore, as it has been noted, the introduction of evidence of mechanisms into EBP faces important challenges. On the one hand, evidence of mechanisms has relevant limitations. Certainly, those problems are less threatening in certain scenarios, and, in some cases, they may be mitigated by complementary evidence of correlations. However, they may constitute a serious threat to the contribution by evidence of mechanisms and should not be overlooked. On the other hand, there are relevant divergences between EBM and EBP concerning both mechanisms and mechanistic studies. In order to integrate evidence of mechanisms in the EBP framework, consequently, it is crucial to develop guidelines for assessing and evaluating this kind of evidence. Those guidelines should take into account the particularities of evidence of mechanisms in EBP.

[2] It should be noted that, as some advocates of ABMs have acknowledged, generative sufficiency is hardly sufficient to conclusively establish the mechanism responsible for a phenomenon (Macy & Flache, 2009). There may be alternative configurations, which may remain unknown, able to produce the same output.

Funding

This work was supported by the Spanish Ministry of Science and Innovation [grant number PID2021-125936NB-I00].

CRediT authorship contribution statement

Saúl Pérez-González: Conceptualization, Funding acquisition, Investigation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

None.

Acknowledgements

I would like to thank Valeriano Iranzo, María Jiménez-Buedo, Jan Sprenger, and two anonymous reviewers for their helpful comments.

References

Anjum, R. L., Copeland, S., & Rocca, E. (2020). Medical scientists and philosophers worldwide appeal to EBM to expand the notion of 'evidence'. *BMJ Evidence-Based Medicine, 25*(1), 6–8. https://doi.org/10.1136/bmjebm-2018-111092

Aronson, J. K., Caze, A. L., Kelly, M. P., Parkkinen, V.-P., & Williamson, J. (2018). The use of mechanistic evidence in drug approval. *Journal of Evaluation in Clinical Practice, 24*(5), 1166–1176. https://doi.org/10.1111/jep.12960

Baron, J. (2018). A brief history of evidence-based policy. *The Annals of the American Academy of Political and Social Science, 678*(1), 40–50. https://doi.org/10.1177/0002716218763128

Barrett, C. B., & Carter, M. R. (2010). The power and pitfalls of experiments in development economics: Some non-random reflections. *Applied Economic Perspectives and Policy, 32*(4), 515–548. https://doi.org/10.1093/aepp/ppq023

Beach, D., & Pedersen, R. B. (2019). *Process-tracing methods: Foundations and guidelines* (2nd ed.). University of Michigan Press.

Biggs, M. (2009). Self-fulfilling prophecies. In P. Hedström, & P. S. Bearman (Eds.), *The oxford handbook of analytical sociology* (pp. 294–314). Oxford University Press.

Bohrnstedt, G. W., & Stecher, B. M. (2002). *What we have learned about class size reduction in California. Capstone report.* California Department of Education.

Cabinet Office. (2013). *What works: Evidence centres for social policy.* Cabinet Office.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*(1), 11–20. https://doi.org/10.1017/S1745855207005029

Cartwright, N. (2009). What are randomised controlled trials good for? *Philosophical Studies, 147*(1), 59–70. https://doi.org/10.1007/s11098-009-9450-2

Cartwright, N. (2012). Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79*(5), 973–989. https://doi.org/10.1086/668041

Cartwright, N., Goldfinch, A., & Howick, J. (2009). Evidence-based policy: Where is our theory of evidence? *Journal of Children's Services, 4*(4), 6–14. https://doi.org/10.5042/jcs.2010.0017

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

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine, 57*(6), 745–747. https://doi.org/10.1016/j.ypmed.2012.10.020

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*(2), 339–360. https://doi.org/10.1007/s11245-013-9220-9

Cohen, J., & Dupas, P. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *Quarterly Journal of Economics, 125*(1), 1–45. https://doi.org/10.1162/qjec.2010.125.1.1

Cowen, N., Virk, B., Mascarenhas-Keyes, S., & Cartwright, N. (2017). Randomized controlled trials: How can we know "what works". *Critical Review, 29*(3), 265–292. https://doi.org/10.1080/08913811.2017.1395223

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Davies, P., & Boruch, R. (2001). The Campbell Collaboration: Does for public policy what Cochrane does for health. *BMJ, 323*(7308), 294–295. https://doi.org/10.1136/bmj.323.7308.294

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature, 48*(2), 424–455. https://doi.org/10.1257/jel.48.2.424

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Demeulenaere, P. (2021). Norms. In G. Manzo (Ed.), *Research handbook on analytical sociology* (pp. 233–248). Edward Elgar Publishing.

Drèze, J. (2018). Evidence, policy and politics: A commentary on Deaton and Cartwright. *Social Science & Medicine, 210*, 45–47. https://doi.org/10.1016/j.socscimed.2018.04.025

van Eersel, G. G., Koppenol-Gonzalez, G. V., & Reiss, J. (2019). Extrapolation of experimental results through analogical reasoning from latent classes. *Philosophy of Science, 86*(2), 219–235. https://doi.org/10.1086/701956

Flache, A., & de Matos Fernandez, C. A. (2021). Agent-based computational models. In G. Manzo (Ed.), *Research handbook on analytical sociology* (pp. 453–473). Edward Elgar Publishing.

French, R. D. (2019). Is it time to give up on evidence-based policy? Four answers. *Policy & Politics, 47*(1), 151–168. https://doi.org/10.1332/030557318X15333033508220

Gamoran, A. (2018). Evidence-based policy in the real world: A cautionary view. *The Annals of the American Academy of Political and Social Science, 678*(1), 180–191. https://doi.org/10.1177/0002716218770138

Gelman, A. (2018). Benefits and limitations of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine, 210*, 48–49. https://doi.org/10.1016/j.socscimed.2018.04.034

Gillies, D. (2017). Evidence of mechanism in the evaluation of streptomycin and thalidomide. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 66*, 55–62. https://doi.org/10.1016/j.shpsc.2017.06.003

Gillies, D. (2018). *Causality, probability, and medicine*. Routledge. https://doi.org/10.4324/9781315735542

Gilmore, E. A., Herzer Risi, L., Tennant, E., & Buhaug, H. (2018). Bridging research and policy on climate change and conflict. *Current Climate Change Reports, 4*(4), 313–319. https://doi.org/10.1007/s40641-018-0119-9

Gilson, L., & McIntyre, D. (2005). Removing user fees for primary care in Africa: The need for careful action. *BMJ, 331*(7519), 762–765. https://doi.org/10.1136/bmj.331.7519.762

Gluckman, P., & Wilsdon, J. (2016). From paradox to principles: Where next for scientific advice to governments? *Palgrave Communications, 2*(1), 1–4. https://doi.org/10.1057/palcomms.2016.77

Grüne-Yanoff, T. (2016). Why behavioural policy needs mechanistic evidence. *Economics and Philosophy, 32*(3), 463–483. https://doi.org/10.1017/S0266267115000425

Haskins, R. (2018). Evidence-based policy: The movement, the goals, the issues, the promise. *The Annals of the American Academy of Political and Social Science, 678*(1), 8–37. https://doi.org/10.1177/0002716218770642

Haynes, L., Goldacre, B., & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. Cabinet Office-Behavioural Insights Team.

Head, B. W. (2008). Three lenses of evidence-based policy. *Australian Journal of Public Administration, 67*(1), 1–11. https://doi.org/10.1111/j.1467-8500.2007.00564.x

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*, 49–67. https://doi.org/10.1146/annurev.soc.012809.102632

Howick, J., Glasziou, P., & Aronson, J. K. (2010). Evidence-based mechanistic reasoning. *Journal of the Royal Society of Medicine, 103*(11), 433–441. https://doi.org/10.1258/jrsm.2010.100146

Howick, J., Glasziou, P., & Aronson, J. K. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical Medicine and Bioethics, 34*(4), 275–291. https://doi.org/10.1007/s11017-013-9266-0

IARC. (2012). *IARC monographs on the evaluation of carcinogenic risks to humans: Vol. 100F. Chemical agents and related occupations*. International Agency for Research on Cancer.

IARC. (2019). *IARC monographs on the identification of carcinogenic hazards to humans: Preamble*. International Agency for Research on Cancer.

Illari, P. M. (2011). Mechanistic evidence: Disambiguating the russo–williamson thesis. *International Studies in the Philosophy of Science, 25*(2), 139–157. https://doi.org/10.1080/02698595.2011.574856

James, C. D., Hanson, K., McPake, B., Balabanova, D., Gwatkin, D., Hopwood, I., et al. (2006). To retain or remove user fees? *Applied Health Economics and Health Policy, 5*(3), 137–153. https://doi.org/10.2165/00148365-200605030-00001

Kemm, J. (2006). The limitations of 'evidence-based' public health. *Journal of Evaluation in Clinical Practice, 12*(3), 319–324. https://doi.org/10.1111/j.1365-2753.2006.00600.x

Khosrowi, D. (2019). Trade-offs between epistemic and moral values in evidence-based policy. *Economics and Philosophy, 35*(1), 49–78. https://doi.org/10.1017/S0266267118000159

Khosrowi, D., & Reiss, J. (2019). Evidence-based policy: The tension between the epistemic and the normative. *Critical Review, 31*(2), 179–197. https://doi.org/10.1080/08913811.2019.1688520

Knox, V., Hill, C. J., & Berlin, G. (2018). Can evidence-based policy ameliorate the nation's social problems? *The Annals of the American Academy of Political and Social Science, 678*(1), 166–179. https://doi.org/10.1177/0002716218769844

La Caze, A., & Colyvan, M. (2017). A challenge for evidence-based policy. *Axiomathes, 27*(1), 1–13. https://doi.org/10.1007/s10516-016-9291-5

Lagarde, M., & Palmer, N. (2011). The impact of user fees on access to health services in low- and middle-income countries. *Cochrane Database of Systematic Reviews, 4*, CD009094. https://doi.org/10.1002/14651858.CD009094

Leuridan, B., & Weber, E. (2011). The IARC and mechanistic evidence. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 91–109). Oxford University Press.

Lisciandra, C. (2020). Reflections on the 2019 Nobel Memorial Prize Awarded to Banerjee, Duflo, and Kremer. *Erasmus Journal for Philosophy and Economics, 13*(1), 79–92. https://doi.org/10.23941/ejpe.v13i1.487

Littell, J. H., & White, H. (2018). The Campbell collaboration: Providing better evidence for a better world. *Research on Social Work Practice, 28*(1), 6–12. https://doi.org/10.1177/1049731517703748

Lohse, S., & Canali, S. (2021). Follow *the* science? On the marginal role of the social sciences in the COVID-19 pandemic. *European Journal for Philosophy of Science, 11*(4), 99. https://doi.org/10.1007/s13194-021-00416-y

Macy, M., & Flache, A. (2009). Social dynamics from the bottom up: Agent-based models of social interaction. In P. Hedström, & P. S. Bearman (Eds.), *The oxford handbook of analytical sociology* (pp. 245–268). Oxford University Press.

Moneta, A., & Russo, F. (2014). Causal models and evidential pluralism in econometrics. *Journal of Economic Methodology, 21*(1), 54–76. https://doi.org/10.1080/1350178X.2014.886473

Montuschi, E. (2009). Questions of evidence in evidence-based policy. *Axiomathes, 19*(4), 425–439. https://doi.org/10.1007/s10516-009-9085-0

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2), 113–127. https://doi.org/10.2307/1602360

Osimani, B. (2013). Until RCT proven? On the asymmetry of evidence requirements for risk assessment. *Journal of Evaluation in Clinical Practice, 19*(3), 454–462. https://doi.org/10.1111/jep.12039

Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., et al. (2018). *Evaluating evidence of mechanisms in medicine: Principles and procedures*. Springer.

Parsons, W. (2002). From muddling through to muddling up—evidence based policy making and the modernisation of British government. *Public Policy and Administration, 17*(3), 43–60. https://doi.org/10.1177/095207670201700304

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review—a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy, 10*(1), 21–34. https://doi.org/10.1258/1355819054308530

Pérez-González, S., & Iranzo, V. (2021). Assessing the role of evidence of mechanisms in causal extrapolation. *THEORIA. An International Journal for Theory, History and Foundations of Science, 36*(2), 211–228. https://doi.org/10.1387/theoria.21642

Reiss, J. (2010). Across the boundaries: Extrapolation in biology and social science, Daniel P. Steel. Oxford University Press, 2007. xi + 241 pages. *Economics and Philosophy, 26*(3), 382–390. https://doi.org/10.1017/S0266267110000325

Robert, E., & Ridde, V. (2013). Global health actors no longer in favor of user fees: A documentary study. *Globalization and Health, 9*(1), 29. https://doi.org/10.1186/1744-8603-9-29

Rocca, E., Anjum, R. L., & Mumford, S. (2020). Causal insights from failure: Post-marketing risk assessment of drugs as a way to uncover causal mechanisms. In A. LaCaze, & B. Osimani (Eds.), *Uncertainty in pharmacology: Epistemology, methods, and decisions* (pp. 39–57). Springer.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*(2), 157–170. https://doi.org/10.1080/02698590701498084

Sakaguchi, K., Varughese, A., & Auld, G. (2017). Climate wars? A systematic review of empirical analyses on the links between climate change and violent conflict. *International Studies Review, 19*(4), 622–645. https://doi.org/10.1093/isr/vix022

Scottish Intercollegiate Guidelines Network. (2011). *Sign 50: A guideline developer's handbook*. Scottish Intercollegiate Guidelines Network. https://www.sign.ac.uk/assets/sign50_2011.pdf.

Shan, Y., & Williamson, J. (2021). Applying evidential pluralism to the social sciences. *European Journal for Philosophy of Science, 11*(4), 96. https://doi.org/10.1007/s13194-021-00415-z

Steel, D. (2004). Social mechanisms and causal inference. *Philosophy of the Social Sciences, 34*(1), 55–78. https://doi.org/10.1177/0048393103260775

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

Tonelli, M. R., & Williamson, J. (2020). Mechanisms in clinical practice: Use and justification. *Medicine, Healthcare & Philosophy, 23*(1), 115–124. https://doi.org/10.1007/s11019-019-09915-5

Vandenbroucke, J. P. (2006). What is the best evidence for determining harms of medical treatment? *Canadian Medical Association Journal, 174*(5), 645–646. https://doi.org/10.1503/cmaj.051484

Vandenbroucke, J. P. (2008). Observational research, randomised trials, and two views of medical science. *PLoS Medicine, 5*(3), e67. https://doi.org/10.1371/journal.pmed.0050067

Victora, C. G., Habicht, J.-P., & Bryce, J. (2004). Evidence-based public health: Moving beyond randomized trials. *American Journal of Public Health, 94*(3), 400–405.

Wagenaar, H. (2007). Governance, complexity, and democratic participation: How citizens and public officials harness the complexities of neighborhood decline. *The American Review of Public Administration, 37*(1), 17–50. https://doi.org/10.1177/0275074006296208

Wilde, M., & Parkkinen, V.-P. (2019). Extrapolation and the russo–williamson thesis. *Synthese, 196*(8), 3251–3262. https://doi.org/10.1007/s11229-017-1573-y

Williamson, J. (2019). Establishing causal claims in medicine. *International Studies in the Philosophy of Science, 32*(1), 33–61. https://doi.org/10.1080/02698595.2019.1630927

Wong, G., Westhorp, G., Pawson, R., & Greenhalgh, T. (2013). *Realist synthesis: RAMESES training materials*. RAMESES Project.

Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass, 2*(6), 981–1022. https://doi.org/10.1111/j.1747-9991.2007.00106.x