Philosophy of Medicine
Original Research

Interrogating Evidential Pluralism: A Critical Comparison of Approaches to Incorporating Mechanisms in Evidence Synthesis

Trisha Greenhalgh[1]

1 Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom.
Email: trish.greenhalgh@phc.ox.ac.uk

Abstract

Different versions of evidential pluralism offer contrasting accounts of how evidence of mechanism should be combined with evidence of association for the purposes of causal inference. This paper explores the philosophical learnings from an empirical application of two approaches—formal Evidential Pluralism (developed relatively recently and associated with Williamson) and traditional evidential pluralism (which has a long history in philosophy of science; its more recent iterations are associated with, among others, Russo and Illari). These approaches are contrasted using a complex and contested empirical case: a heterogeneous body of studies on face mask efficacy, offering evidence of both association and mechanism. The findings raise questions about the adequacy of formal Evidential Pluralism for questions of causal inference in complex healthcare contexts.

1. Introduction

1.1 Outline of Argument

How should evidence of mechanisms be integrated with evidence of association when assessing causality in medicine? This question is central to the debate around evidential pluralism, which is sometimes incorrectly depicted as originating in an influential paper which introduced the Russo-Williamson thesis (RWT) (Russo and Williamson 2007). The RWT argues that establishing a causal claim requires both evidence of *probabilistic dependencies* (typically from randomized controlled trials [RCTs] or observational studies) and evidence of *physical mechanisms* (typically from basic science). The term "probabilistic dependencies" has largely been replaced in more recent studies with "association" or "difference-making"—that is, that a change in one variable (the putative cause) is associated with a change in another variable (the putative effect). Citing work by Wesley C. Salmon (1998) and Phil Dowe (2000), Federica Russo and Jon Williamson use the term "physical mechanisms" (2007, 164) to refer primarily to physical processes and interactions that can be explained by the "laws of nature," but they also suggest that the term could embrace

Pitt Open Library Publishing (cc) BY

This work is published by Pitt Open Library Publishing and is licensed under a Creative Commons Attribution 4.0 International License. © The Author(s).

ISSN 2692-3963 | Date Received 20-05-2026 | Date Accepted 28-08-2026 | Date First Published 01-10-2026
Philosophy of Medicine | DOI 10.5195/pom.2026.316 | Volume 7 | Issue 1 | pp. 1–26

processes in the social sciences (that is, social and political mechanisms). In that sense, they distinguish a "physical" mechanism from a purely statistical relationship.

Different versions of evidential pluralism offer contrasting and, at times, incompatible accounts of how different kinds of evidence relate to one another. These reflect disagreements about the nature of causal inference, including whether mechanisms and associations can be assessed independently of each other, how confounding and bias should be analyzed in practice, and what role formal frameworks can play in guiding empirical inquiry.

This paper explores the philosophical learnings from an empirical evaluation of two approaches to integrating evidence of association with evidence of mechanism—Evidential Pluralism (associated with Williamson) and traditional evidential pluralism (associated with, among many others, Phyllis McKay Illari and Russo).[1] These approaches are examined using a heterogeneous dataset of primary studies on face mask efficacy comprising hundreds of papers, including randomized trials, observational studies, and multiple kinds of mechanistic evidence.

This paper should be read as a piece of translational epistemology rather than a systematic review or a purely conceptual comparison. By this I mean philosophical work that takes an account of causal inference out of the seminar room and tries to *use* it: not to illustrate it with examples chosen because they fit but to conduct a real evidence synthesis with it and find out what it can and cannot do. What such an exercise yields is not empirical findings about masks but instead findings about the account. The empirical analyses are reported in full in published papers (Greenhalgh et al. 2026a, 2026b) and in preparation; what is offered here is a philosophical account of what happened when two approaches for combining evidence of mechanism with evidence of association were applied in earnest to a complex body of evidence. Empirical detail is provided only insofar as it is needed to show where each approach held and where it gave way.

For the purposes of this paper, the question "what is the efficacy of mask-wearing on transmission of respiratory infections?" was used as an empirical stress test for different philosophical approaches to causal inference. The case provided an unusually rich setting in which to examine how different approaches to evidential pluralism coped with causal complexity, dynamic contexts, and multiple interacting sources of bias.

When applied to the complex dataset on mask efficacy, Evidential Pluralism provided a clear and elegant conceptual framework but proved difficult to operationalize beyond a heuristic level. It appeared to presuppose relatively simple and stable causal structures, and it treated mechanistic and associative evidence as unproblematically separable. Furthermore, while the framework could in principle accommodate confounding and bias, it offered no procedure for analyzing these phenomena in a large, heterogeneous epidemiological dataset characterized by multiple interacting sources of distortion.

By contrast, the traditional evidential pluralism literature, whose historical roots pre-date the 2007 RWT paper by many decades, offered methodological tools better able to handle these challenges. Early papers by Illari and Russo (separately and together), along with subsequent developments of this broader traditional framing, informed an empirical

[1] Williamson capitalizes the term "Evidential Pluralism" to denote a specific, formalized methodology, described later in this paper. To distinguish this version, I will use the notation "Evidential Pluralism." In this paper, the uncapitalized term "evidential pluralism" refers to the older and broader tradition, which is not associated with a single individual and does not recommend specific steps.

analysis that reflected (i) the complexity and distributed nature of causal mechanisms, (ii) the empirical inseparability of mechanistic and associative evidence, and (iii) the need to analyze confounding and other sources of bias alongside mechanistic reasoning.

The paper concludes by hypothesizing that Evidential Pluralism is too rigid in its sequence and too indeterminate in its guidance to support causal inference in complex healthcare contexts. A more adequate approach is needed that treats mechanistic and associative evidence as inseparable and integrates them, as appropriate, with a systematic quantitative analysis of confounding and other sources of bias. This has implications for how evidence is conceptualized and synthesized in healthcare, particularly in relation to the interpretation of heterogeneous and apparently conflicting findings from epidemiological studies.

The paper makes three main contributions. First, it offers a critique of Evidential Pluralism, grounded in its limited ability to guide analysis in a complex empirical domain. Second, it mobilizes the older and broader version of evidential pluralism, showing that its treatment of mechanisms and associations as intertwined, and its insistence that confounding be addressed alongside mechanistic reasoning, are better suited to heterogeneous epidemiological datasets. Third, it argues that neither approach has yet supplied what such datasets require: an account of biases and confounders as mechanisms in their own right, with direction and magnitude, and therefore as proper objects of the same evidential appraisal as the mechanisms through which an intervention is supposed to act.

More broadly, the paper illustrates how difficulties encountered in empirical research can reveal limitations in the applicability of existing philosophical approaches and point toward alternatives that are more fit for purpose.

1.2 Evidential Pluralism and Evidence-Based Medicine: A Brief History

Philosophers of science have long argued that robust causal inference requires integrating multiple kinds of evidence from different methods and sources—see, for example, work by Nancy Cartwright (1983), Donald Campbell (1987), Wesley C. Salmon (1998), and Sandra Harding and Merrill Hintikka (2003). This broad tradition of evidential and methodological pluralism spans diverse intellectual traditions. I do not attempt to review that literature here. Rather, I begin with the RWT, partly because it can be seen as a particularly explicit and influential formulation of a much older idea: that causal claims in science are best supported by converging lines of evidence of different kinds, and partly because it focuses specifically on the relationship between two kinds of evidence—that of mechanism and that of association.

In particular, the RWT has contributed significantly to efforts by philosophers of science to shift evidence-based medicine (EBM) beyond the conceptual constraints of its "hierarchy of evidence" as laid out, for example, in early EBM textbooks (Sackett et al. 1991) and widely reproduced since. In this hierarchy, RCTs are viewed as providing more reliable evidence than other study designs, notably non-randomized observational studies and studies that demonstrate mechanisms (for philosophical critiques of this hierarchy of evidence, see Borgerson [2009], Goldenberg [2009], Clarke et al. [2014], and Stegenga [2014]).

Russo and Williamson's (2007) landmark paper, along with wider work arguing that mechanisms and associations should be seen as complementary and mutually reinforcing

(Clarke et al. 2014, 2013; Gillies 2017), influenced the emergence of EBM+ (evidence-based medicine plus). As Veli-Pekka Parkkinen et al. (2018) describe in their book, the EBM+ Consortium was formed in about 2017 by a group of philosophers, epidemiologists, statisticians, and clinicians who sought to incorporate mechanistic evidence more systematically into medical reasoning. Several members of that Consortium had links to the EBM movement and sought to challenge *from within* the low status of mechanistic evidence in its hierarchy of evidence: "The methods in this book build upon the tools already developed by EBM, by taking evidence of mechanisms into account in addition to the clinical studies that are the focus of EBM. We refer to this development as EBM+" (Parkkinen et al. 2018, 3).

The EBM+ Consortium campaigned for EBM scholars to treat mechanistic evidence as *complementary* rather than *inferior* to associative evidence, to make all evidence (that is, of both association *and* mechanism) explicit and to evaluate all this evidence systematically. This represented a change from conventional EBM, which took some mechanistic evidence (for example, upstream studies in early drug development) as given but did not advocate for its inclusion or formal appraisal when making policy or treatment decisions (Guyatt et al. 1992).

EBM+ was not without its philosophical critics. Jeremy Howick (2011) raised concerns that mechanisms are (sometimes if not always) "just-so stories," that is, plausible-sounding but weakly supported narratives. Bennett Holman (2019) cited examples to show that plausible mechanisms (for example, in relation to so-called anti-arrhythmic drugs, which turned out to *cause* fatal arrhythmias) had sometimes misled doctors, with tragic consequences for patients. But others argued that EBM+ was offering a systematic and philosophically rigorous way of *testing* mechanisms alongside associative evidence (see, for example, Williamson [2021b]).

While the various scholars in the EBM+ project had much in common (not least, a commitment to challenging EBM's tendency to dismiss non-RCT evidence), there were important disagreements among them. As I argue below, one crucial bone of contention was whether it is helpful to talk about *mechanistic (or associative) evidence* and *mechanistic (or associative) studies* as opposed to *evidence of mechanisms (or association)*. These framings differ in the extent to which different kinds of empirical evidence can be separated into "associative" and "mechanistic" and analyzed as such.

In recent years, EBM+ has become closely associated with the Evidential Pluralism approach developed by Williamson and described below. This version of EBM+ is increasingly popular as an idea in medical circles (see, for example, Aronson et al. [2018], Abdin et al. [2019], Auker-Howlett and Williamson [2021], Greenhalgh et al. [2022], Pfaff and Schmitt [2023]). But, at the time of writing, a detailed, empirically grounded methodology for *how* to combine evidence of association with evidence of mechanism in EBM+ has not been published.

2. The Empirical Case

2.1 The Research Study

This paper emerged during a research study, Interdisciplinary Systematic Review: Mechanistic Evidence and Epistemic Justice, funded by UK Research and Innovation, on

which Jon Williamson was chief investigator, and I was a co-investigator; other team members are listed in the Acknowledgments. The study combined a philosophical analysis of causal inference with empirical examples, the main one being the efficacy of masks and mask mandates in reducing transmission of respiratory disease.

The question that concerned us was how to combine associative evidence of mask efficacy (from RCTs and observational studies) and mask mandate efficacy (for example, from ecological and quasi-experimental studies of policy responses) with mechanistic evidence of *how* these interventions might work. Various empirical papers are published or under review. This paper focuses on philosophical approaches as applied to the first part of the systematic review (efficacy of masks in the community); work on datasets relating to masks in healthcare settings and mask mandates is ongoing.

2.2 How the Approaches Were Tested

Any study in the philosophy of science or medicine combines a doing component and a thinking component. In this case, the doing component was desk research—developing and applying a systematic methodology for synthesizing primary research studies. While no fewer than 66 teams had already attempted to synthesize evidence on mask efficacy (Greenhalgh et al. 2026b), we were the first explicitly to seek out mechanistic evidence and combine it with the (better-known and often more highly valued) evidence from RCTs and other associative studies. Our review involved a thorough search for potentially relevant papers, sifting and sorting to hone the dataset, critical close reading of empirical studies, and synthesis of findings.

In terms of the thinking component (the challenge of combining different kinds of evidence), a key element was the role of doubt and disagreement. Paul K.J. Han defines uncertainty as "the attitude of committed neutrality that opens our eyes to our ignorance" (2021, 21). My work for this paper was driven by my growing uneasiness that the methods we had begun with, as set out in the grant application and the protocol paper (Greenhalgh et al. 2025), which talked about extending Evidential Pluralism to systematic review in an approach we called "SR+," were not working well for the mask efficacy dataset.

Different views among the team generated some healthy debates as well as some conflict. Partway through the project, we decided to pilot two contrasting approaches: the original multistep SR+ method, led by Williamson, and a more interpretive approach, mechanism-informed narrative synthesis (MINS), led by me. The latter method has been applied to a non-mask empirical case (Helm et al. 2026); a paper describing its detailed application to the mask efficacy dataset is under review at the time of writing.

2.3 The Mask Efficacy Dataset

The question of face mask efficacy in reducing respiratory disease transmission is broad and complex. As it stands, it covers all kinds of respiratory diseases, all kinds of face covering, all populations, all ways of incentivizing and encouraging people to wear masks, all settings, and all of history. A systematic search identified around 400 papers reporting studies testing the *association* between mask-wearing (and advice to mask) and some disease-related outcome. Only 26 of these were RCTs; around 150 were observational studies (for example, cohort, case control, cross-sectional survey) and most of the remainder were quasi-experimental, ecological, and mathematical modeling studies, including natural

experiments of mask mandates. In addition, numerous mechanistic studies addressed elements on the causal pathway directly linking masks and disease, while many more addressed related pathways. Box 1 describes examples of these studies.

Box 1: Examples of Primary Studies of Mask Efficacy and Relevant Mechanistic Evidence

Example 1: Cowling et al. (2009)
This RCT from Hong Kong, cluster-randomized by household, evaluated masking and hand hygiene to reduce secondary infection of family members by an index case with confirmed seasonal influenza. 407 index patients with 794 household contacts were recruited from fever clinics and randomized to masks + hand hygiene, hand hygiene alone, or control arm. Self-reported adherence to masking was poor (49% of index cases and 26% of household contacts said they masked) and level of missing data was high (more than a third of people randomized had no follow-up data). There was no overall difference in secondary attack rate between arms but a prespecified early-implementation subgroup, in which masking commenced within 36 hours of the index case developing symptoms, showed masks + hand hygiene was significantly more effective than control (adjusted odds ratio 0.33 [95% CI 0.13–0.87]) for preventing PCR-confirmed influenza (the hand hygiene alone arm had adjusted odds ratio 0.46, which was not statistically significant). The prespecified subgroup analysis is biologically plausible as it covers the period of maximum shedding in the early symptomatic phase (see example 2).

Example 2: Suess et al. (2010)
This German study collected details of symptoms and also swabs for viral culture every day for 10 days from 83 people in 36 households during an outbreak of H1N1 influenza virus. They showed that viral shedding was highest in the first two days of symptoms and fell substantially thereafter.

Example 3: Alfelali et al. (2020)
This RCT from Saudi Arabia randomized more than 7,500 participants over 3 years. All were pilgrims attending the annual hajj—a mass religious event, which involves sleeping in communal tents for several days and performing a series of rituals, some in indoor crowded spaces. Cluster-randomization was by tent; medical student assistants visited the tents twice daily to inquire about symptoms and (for the intervention arm) distribute masks and encourage masking. Symptomatic participants were offered the gold standard PCR test for viral respiratory diseases, including the common cold and influenza. The study showed no significant protective effect of masks by either intention-to-treat or per-protocol analysis. This null finding was consistent with lack of efficacy but was also consistent with powerful signal attenuation mechanisms linked to airborne transmission (see example 5), including shared sleeping arrangements (tent occupancy averaged 75 people and pilgrims slept head-to-head; nobody masked overnight), poor adherence, and crossover (24.7% and 14.3% of people in the intervention and control arms, respectively, wore masks most of the time in the daytime; 47.7% and 34.9% wore them intermittently).

Example 4: Lee et al. (2020)
This study from South Korea looked at more than 300 people infected with SARS-CoV-2, more than one-third of whom were asymptomatic. Using serial PCR measurements, they estimated viral shedding over several days. Asymptomatic patients shed similar viral loads to symptomatic ones.

Example 5: Kutter et al. (2021)
In a laboratory experiment from a Dutch team, a ferret infected with a virus (influenza, SARS or SARS-CoV-2) was placed in a sealed cage. The cage was connected by air ducts to another cage containing a healthy ferret. The duct system included several 90º angles to exclude droplet transmission. In 25 of 26 runs of the experiment, the healthy ferret subsequently became infected with the same virus. The authors concluded that these viruses traveled through the air and remained viable as they did so.

Example 6: Doung-Ngern et al. (2020)
This event-anchored case control study from Thailand during the first wave of the Covid-19 pandemic focused on contacts of cases from 3 major clusters. There was PCR verification of cases and multilevel adjustment for distance (how far the contact kept from the index case), exposure duration (how long they spent with them), and other behaviors. Of 1,050 asymptomatic contacts, 211 tested positive and 839 negative. Mask wearing during exposure to the infected person was strongly associated with reduced risk of infection. "Always" wearing a mask had an adjusted odds ratio of 0.23 (95% CI 0.09–0.6), while "sometimes" wearing a mask was not protective. Masking, duration of contact, and distancing all had independent effects, consistent with airborne transmission mechanisms. Findings were confounded by other protective behaviors: those who "always" masked also tended to maintain distance, have shorter contact duration, and wash their hands more; those who masked "sometimes" had longer, closer contacts. While the authors adjusted for these variables, they relied on a stepwise, data-driven selection process, increasing the chance that subtle confounding remained. Selection bias was likely, since inclusion was conditional on being identified as a contact and reachable by phone. The lack of protective effect observed for "sometimes" masking is consistent with an airborne transmission route (see example 5) and intermittent use in high-risk settings, where brief periods of intense exposure may overwhelm partial filtration.

Example 7: Schmitt and Wang (2022)
This review and computer modeling study synthesized 12 studies of mask leakage using mannikins and artificial air currents in the laboratory. Leakage around the edges of a mask or respirator (due to poor fit) was consistently shown to be a more important contributor to mask failure than poor filtration. In the laboratory setting, respirators consistently fitted better than medical masks (and medical masks better than cloth masks), thereby directing a higher proportion of air through the filtering part.

The studies testing association were highly heterogeneous (see examples 1, 3, and 6 in Box 1). Their designs were typically pragmatic, reflecting the realities of a rapidly spreading infectious disease outbreak in a unique setting. Measures of intervention (for example, whether and when someone masked) and outcome (whether disease transmission had occurred) were inconsistent across studies. Sources of bias (for example, low adherence, differential loss to follow-up) and confounding (for example, demographic risk factors, covarying preventive behaviors) were ubiquitous.[2]

[2] Bias refers to any systematic error in the design, conduct, analysis or reporting of a study that leads to a distorted estimate of the association between an exposure and an outcome. Confounding occurs when the observed association between an exposure and an outcome is distorted by a third variable that is causally related

Mechanistic evidence directly or indirectly relevant to how face masks might work comes from multiple disciplines. It includes, for example, the pathophysiology of disease and its transmission, the physics of particle transmission in fluids, the material science of mask fabrics, the aerodynamics of mask leakage, the mathematics of epidemic dynamics, the psychology of human behavior, and more.

Given these features of the primary evidence, I found that making sense of it required an iterative, detailed and non-straightforward process, which included:

a) dividing the dataset pragmatically into manageable subsets (for example, mass gathering studies, occupational health studies);
b) within each subset, identifying individual studies that tested an association between masking (or advice to mask) and disease;
c) appraising these studies carefully using epidemiological tools for estimating confounding and other sources of bias;
d) questioning whether and to what extent each study had been designed to take account of key mechanisms such as viral shedding (see examples 2 and 4 in Box 1 [Suess et al. 2010; Lee et al. 2020]), airborne transmission (example 5 in Box 1 [Kutter et al. 2021]), and the distorting influence of mask leakage (example 7 in Box 1 [Schmitt and Wang 2022]);
e) incorporating relevant wider scientific knowledge—for example, incubation periods, overdispersal (where, in some diseases, a small number of individuals account for a large number of secondary cases), and the behavioral science of public health interventions;
f) examining heterogeneous findings (for example, why some studies demonstrated statistically significant effects while others did not); and
g) progressively synthesizing findings into a nuanced and coherent account that explained all relevant evidence.

Absent from the above account is the precise method for integrating evidence of association with evidence of mechanism. As I recount below, developing that method was challenging. In the remaining sections, I describe how each version of evidential pluralism performed as I tried to apply it in a way that was both rationally grounded and empirically defensible.

3. Applying Evidential Pluralism

3.1 The Diagram and Steps

> Evidential Pluralism is a theory of causal enquiry which maintains that one needs to scrutinise mechanistic studies alongside the experimental and observational studies that are the focus of present-day evidence-based medicine and evidence-based policy. (Trofimov and Williamson 2025, 647)

Evidential Pluralism is best known as a diagram (Figure 1). The diagram was originally published in a 2018 blog by Williamson and later reproduced in Chapter 2 (by Williamson)

to both but is not on the causal pathway between them. Confounding reflects underlying causal structures that generate associations between variables; bias arises when such structures (and other sources of systematic error) are not adequately accounted for in study design or analysis. In this paper, I treat both bias and confounding as phenomena that can be analyzed mechanistically, since each arises from underlying causal processes that shape observed associations.

in *Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures* (Parkkinen et al. 2018) as well as in multiple subsequent publications by Williamson and others (see below).

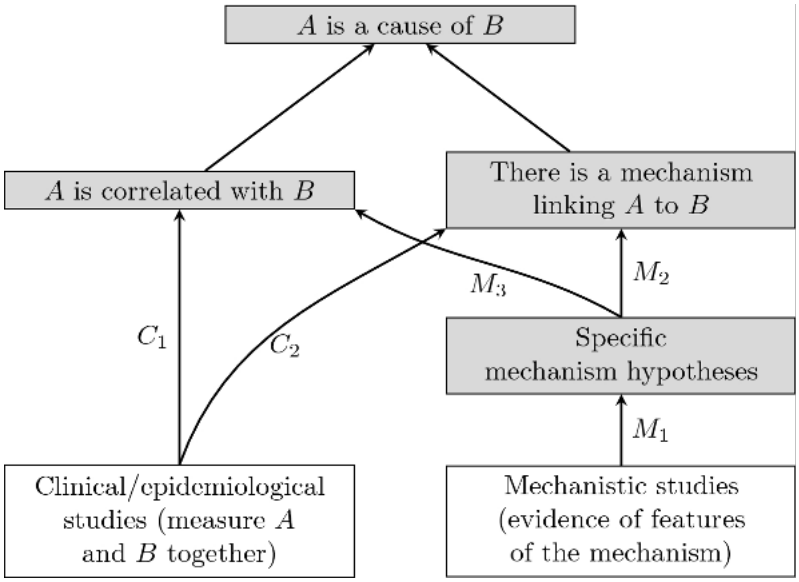


Figure 1. Williamson's widely cited Evidential Pluralism diagram
Reproduced under Creative Commons Attribution 4.0 license from Williamson (2021a)

The diagram is, arguably, a good point at which to begin the epistemological journey from the "hierarchy of evidence" beloved of traditional EBM scholars (Sackett et al. 1991), which ranks studies that test association above studies that elucidate mechanisms, to a position that views evidence of association and evidence of mechanism as complementary and mutually reinforcing in demonstrating causality (Clarke et al. 2014). An important element of the diagram is what Williamson calls "channels"—the connecting arrows in Figure 1. These represent examples of when correlational studies confirm the findings of mechanistic ones (arrows C_1 and C_2) or vice versa (arrows M_1, M_2, and M_3).

Conceptually, an association (for example, that people who smoke are many times more likely to get lung cancer than those who do not) is very different from a mechanism (for example, that various compounds in tobacco smoke damage DNA, cause chronic inflammation and oxidative stress, and overwhelm repair systems, all of which predispose to cancerous change). But just because associations and mechanisms can be so cleanly separated *conceptually* does not mean that they can be unproblematically separated *empirically*.

Williamson sets out the methodology of applying Evidential Pluralism in a series of steps, beginning with specifying the causal claim(s). The steps are broadly as follows (see, for example, Trofimov and Williamson [2025]). For each claim, specify the associative claim and formulate specific mechanism hypotheses. Search for and assess correlation (association) studies. Screen the need for a mechanistic evaluation. Search for and assess mechanistic studies. Assess specific mechanism hypotheses. Assess specific correlation hypotheses. Assess the general mechanism claim. Assess the causal claim. Harmonize the

results. These steps, to be performed in a particular order, reflect a logical argument organized around a "problem claim," an "intervention claim," and a "benefit claim" that can be represented algebraically (in the language of formal logic) or diagrammatically (with causal arrows connecting putative causes with observed events via mediating mechanisms, as in Figure 1).

3.2 Proof of Concept in the Published Literature

Various papers on Evidential Pluralism reproduce the diagram in Figure 1 and reiterate a version of the logical steps (see, for example, Williamson [2021a]; Wilde [2023]; Shan and Williamson [2021]; Trofimov and Williamson [2025]; Greenhalgh et al. [2022]). But in all these papers—my own included—empirical examples are employed *descriptively and superficially to demonstrate proof of concept,* not analyzed in depth.

Alexandra Trofimov and Williamson (2025), for example, use mask mandates in Covid-19 as an example of a policy question that *could be* informed by Evidential Pluralism. They work through the stepwise sequence, illustrating each step with three or four empirical studies, which are described and evaluated very briefly—for example, "strongly supports the second mechanism hypothesis" (2025, 673). They acknowledge that a full analysis of mask and mask mandate efficacy would require a systematic approach to collecting, evaluating, and synthesizing both associative and mechanistic studies. Williamson, I, and others subsequently obtained a research grant to undertake this task.

For a proof of concept analysis, only one or two studies demonstrating a positive association are needed (adding the caveat "conditional on potential confounders" [Trofimov and Williamson 2025, 652]), along with examples of studies demonstrating a mechanism that could account for it. But in attempting to apply Evidential Pluralism to the full mask efficacy dataset, the framework's clear and logically structured sequence of steps were difficult to sustain in the face of the complexity and heterogeneity of the empirical evidence. I list some specific challenges below.

3.3 Asymmetry: Mechanisms and Null Findings

The framework asks whether an observed positive association (say, that transmission decreased when people wore masks) is causal, and looks to evidence of mechanism to help settle the question. That is a reasonable question, and, in many domains, it is the right one. It was not the primary question this dataset posed. Most of the mask trials had returned null findings, and what the mechanistic evidence told us was not that masks work despite those findings. It told us that those trials *could not have detected an effect if there had been one.* Trial participants co-sleeping unmasked, participants crossing over between arms, masks leaking at the face seal—all this is mechanistic knowledge, but it bears on what the design was capable of measuring rather than on the causal relation.

A null finding may have two explanations: that no causal relationship exists, or that one does exist, but the study's design bypassed the mechanism. Published versions of Evidential Pluralism do allow mechanistic evidence to bear on null findings but in one direction only: A mechanism hypothesis may be invoked to explain away an anomalous result that conflicts with an emerging positive correlation claim. I have been unable to find an account of the converse inference—the one our dataset demanded.

The hajj trial (example 3 in Box 1 [Alfelali et al. 2020]) illustrates the difficulty. It randomized more than 7,500 pilgrims by tent over three years and found no protective effect by either intention-to-treat or per-protocol analysis. The framework establishes that no correlation has been demonstrated and leaves the causal claim unestablished. That conclusion is not wrong, but it is empty, and the mechanistic evidence that would make it informative has nowhere to attach. The mechanisms that matter are the ones that determined what this trial could see. Tents held an average of 75 people sleeping head-to-head, and nobody masked overnight, in a setting where airborne transmission is well evidenced (see example 5 in Box 1 [Kutter et al. 2021]). The exposure most likely to be decisive therefore occurred where the intervention was absent from both arms.

The early paper by Russo and Williamson stated that "probabilistic evidence *needs to be accounted for* by an underlying mechanism before the causal claim can be established" (2007, 159; emphasis added). This may be true, but it is only half the story. A null finding also needs mechanistic explaining. A framework oriented solely to positive associations returns the same verdict on a trial that looked and found nothing as on one that *could not* have found anything.

3.4 Causal Complexity

The framework seemed to presuppose relatively simple and stable causal relationships, typically represented as a single cause operating through one or a small number of mechanisms. By contrast, the empirical studies examined in our review involved multiple interacting variables across different levels (for example, many diseases, populations, kinds of mask, ancillary behavioral support, and environmental contexts), with causal relationships that were context-sensitive and dynamic, since several variables (for example, disease prevalence, disease transmissibility, mask availability, public trust, concurrent policies) changed over time. While it would, in principle, be possible to extend the framework to accommodate such complexity, doing so would risk losing the clarity and tractability that are its main strengths.

3.5 Confounding and Bias

Although Evidential Pluralism acknowledges the importance of confounding and bias, its formal methodological architecture offers limited practical guidance for how these phenomena are to be analyzed in complex epidemiological datasets. In domains such as infectious disease epidemiology, confounding and bias are not peripheral complications but constitutive features of causal inference itself. The interpretation of positive, null, and heterogeneous findings depends critically on understanding how multiple interacting sources of distortion may amplify, attenuate, conceal, or even generate observed associations.

In principle, the Evidential Pluralism framework can accommodate such considerations, since evaluating a proposed causal mechanism necessarily involves assessing alternative explanations for the observed association. On a Bayesian reading especially, confidence in a causal hypothesis should increase as competing confounding explanations are ruled out. But the framework, as currently articulated, provides little

indication of how such analyses should actually be undertaken in practice when confounding structures are complex, dynamic, and only partially observable.

This limitation became particularly salient in the mask efficacy dataset. RCTs were sparse, heterogeneous, and vulnerable to substantial post-randomization biases, including low adherence, crossover, exposure misalignment, and differential loss to follow-up (Greenhalgh et al. 2026a). Observational studies were affected by multiple overlapping behavioral, environmental, and epidemiological confounders (see the example in section 4.1).

Determining the informativeness of such studies required detailed epidemiological analysis, including quantitative bias analysis, temporally structured causal reasoning, and explicit modeling of mechanisms of bias. These analytic tasks did not emerge naturally from the Evidential Pluralism workflow itself but had to be developed through additional methodological work external to the framework.

3.6 Steps That Cannot Be Taken in Order

The difficulty in Evidential Pluralism is not that any particular step is poorly specified, but that the steps cannot be completed in the order given, as illustrated by the hajj trial described in section 3.3. The framework begins by requiring that the associative claim be specified and the relevant mechanism hypotheses formulated. Nothing available at that point identifies which mechanisms matter here. The mechanisms of bias that turn out to be decisive—co-sleeping, limited uptake, crossover—become salient only once the null result is in hand and one asks what could have produced it. The output of the later steps is thus a precondition of the first. This is not a defect that more careful sequencing or more precise wording could repair; it is a consequence of treating a body of evidence whose parts are mutually informing as though it could be processed in a single pass.

3.7 Looking for an Empirical Application

While Evidential Pluralism has been illustrated through various proof-of-concept examples, I was unable to identify any published applications in which the framework had been used prospectively to conduct an in-depth analysis of a large, heterogeneous empirical dataset. One paper by Williamson (2021a) looked promising: "Establishing the Teratogenicity of Zika and Evaluating Causal Criteria" but, on closer reading, it did not seem that Evidential Pluralism had actually been applied. Zika's teratogenicity had *previously* been demonstrated by experts in the field, in three separate review papers (Rasmussen et al. 2016; Krauer et al. 2017; Frank et al. 2016). Those reviews had drawn in turn on three sets of criteria for causality—the generic Bradford Hill criteria (Bradford Hill 1965), Thomas H. Shepard's (1994) criteria, which address whether a drug is teratogenic, and a bespoke set of criteria, adapted from Bradford Hill's by Fabienne Krauer et al. (2017), to inform a systematic review by the WHO Zika Causality Working Group.

The Williamson paper does not reanalyze the 72 primary studies relating to congenital Zika in the systematic review by Krauer et al. (2017), nor does it evaluate the plausibility of their evidence assessments. In relation to biological plausibility, for example, Krauer et al. document viable viral particles isolated from fetal brains following pregnancy termination after maternal Zika infection, a well-demarcated biochemical pathway affecting particular receptors, and replication of Zika in neural tissue *in vitro*. The Williamson paper does not

comment on the input of a consensus panel of experts who deliberated on that evidence. Rather, it maps the three sets of *criteria* (Bradford Hill, Shepard, Krauer) to the Evidential Pluralism diagram reproduced in Figure 1 (for example, "plausibility and coherence [in the Bradford Hill criteria] … assess channel M_2" and "animal experiments and analogy [in the Krauer criteria] … concern C_1 and C_2" [Williamson 2021a, 2512]). On the basis of what looks suspiciously like a crude numerical count (more criteria in all three sets relate to correlation than to mechanisms), the paper concludes that "none of the three sets of criteria offer more than a very rudimentary treatment of the M channels" (2513).

Williamson's paper finishes by recommending that further work be done using Evidential Pluralism to address the Zika question, offering three ways of approaching this. First, by following more specific criteria, such as the characteristics of carcinogenicity set out by the International Agency for Research on Cancer (IARC) (Smith et al. 2016), though in fact the criteria used by Krauer et al. (2017) are more detailed and Zika-specific than the ten mechanisms suggested by the IARC, of which only one—"alters cell proliferation, cell death, or nutrient supply" (2015, 716)—is relevant to this example. Second, "qualitatively," which appears to mean making yes/no judgments about whether a mechanism is supported by evidence using a set of heuristics set out in Parkkinen et al. (2018). Third, quantitatively, using a Bayesian probabilistic approach (a preliminary outline for such an analysis is sketched, but the analysis is not conducted).

The other stream of literature mentioned in Williamson's papers on Evidential Pluralism is the work by the IARC on carcinogenicity (see, for example, Riboli et al. [2023]), in which mechanistic evidence has long been recognized to play an important role. The IARC's methodology, developed independently and continually evolving, is described by Williamson as "very close to that of EBM+" (2021b, 198), and he considers that it "can be conceptualised" in terms of Evidential Pluralism (2019, 960), and that it "conforms closely to Evidential Pluralism" (Trofimov and Williamson 2025, 662). But these claims appear to be drawing broad conceptual parallels rather than demonstrating the empirical purchase of the Evidential Pluralism method.

3.8 Expertise as a Downstream Check

A further difficulty is the issue of where expertise enters. Evidential Pluralism is designed to deliver determinate judgments through a procedure, and none of its ten steps is an expert-evaluation stage. Expertise, when it appears, does so after the analysis is completed: Experts are invited to check it. This is not an oversight but a consequence of the design, since a procedure that required expert judgment at each step would not be a procedure in the sense intended.

The Zika paper (Williamson 2021a) is illustrative. As noted above, it maps three sets of criteria onto the diagram without engaging the deliberations of the expert consensus panel whose judgments those criteria had been used to reach—though it was the panel's deliberation, not the criteria, that settled the matter.

This became a practical difficulty in our own work. The doubt described in section 2.2 reflected, inter alia, a recognition that appraising complex epidemiological studies required expertise beyond that of the core team. As the review progressed, it became necessary to bring in methodologists—in mathematical modeling, causal inference in observational epidemiology, and infectious disease dynamics—not merely to check our study appraisals

but to *orient us to what we should be noticing,* particularly in identifying and exploring sources of bias. Nothing in the original workflow anticipated this. There was no point at which the framework asked what we did not know, or whose judgment was needed to establish it.

In sum, Evidential Pluralism seemed poorly suited to the mask efficacy project because (a) it assumed simple, stable causal relationships; (b) it was asymmetric, directly addressing positive but not null associative findings, and so offered no leverage for challenging the design of studies whose null results it took at face value; (c) it provided limited methodological guidance for analyzing confounding and bias in a complex and heterogeneous epidemiological dataset; (d) its stepwise sequence could not be followed in order, since the later steps produced what the first ones required; (e) no published examples appeared to exist of its detailed application to a large, heterogeneous dataset; and (f) it positioned topic expertise as a check on a completed analysis rather than as constitutive of appraisal.

It is important to point out that work by Williamson is ongoing, aimed at producing a detailed worked example of Evidential Pluralism applied to a different empirical dataset. It remains to be seen whether such work will resolve the tensions identified in this paper between the formal, stepwise architecture of Evidential Pluralism and the iterative, interpretive and confounding-sensitive practices required for analyzing complex epidemiological evidence.

4. Applying Traditional Evidential Pluralism

As noted above, a long tradition in philosophy of science has argued for a plurality of concepts and methods (Cartwright 1983; Campbell 1987; Salmon 1998; Harding and Hintikka 2003). This kind of mixed-methods research is widely demanded by academic, funding, and policy stakeholders across the social sciences—particularly in public health, education, and sociology. It is from this broad and varied tradition that the original term "evidential pluralism," first coined (I believe) by Russo (2006), emerged. In its traditional uncapitalized form, evidential pluralism captures a range of approaches adopted by a linked group of philosophers whose work on mechanisms pre-dated, and in some cases fed into, the EBM+ project described by Parkkinen et al. (2018). Notably, Cartwright (2021) proposes retaining the term "evidential pluralism" for this traditional version and using "evidential *dualism*" for Williamson's narrower adaptation of it.

Elements of this wider stream of work proved extremely helpful when attempting to deal with complexity and disentangle the effects of confounding in the mask efficacy studies (see examples 1, 3, and 6 in Box 1). I list some specific contributions below.

4.1 Confounding

In a paper critiquing the Russo-Williamson thesis, Illari (2011) observes that an association between C and E may be explained by three things: C causes E, chance, and confounding. Identifying a mechanism strengthens the first of these explanations but does not prove it. More needs to be done to disambiguate the three explanations.

One key approach, argues Illari, is to build a rich understanding of the wider science that has a bearing on the C–E association, including mechanisms *other* than those directly linking C and E: "In a well-understood domain, you have a much better idea of the variables that might be confounders—variables that might themselves cause the effect of interest, while also being correlated with the postulated cause" (Illari 2011, 149). She maintains that unless the domain is well understood, and especially if there are multiple interacting variables involved, other causal pathways might produce spurious associations between C and E.

This was exactly the problem we seemed to be facing. The dataset included a handful of RCTs and dozens of observational studies, some of which had positive findings (perhaps genuine, but perhaps spuriously inflated by away-from-null biases). Building on Illari's argument, I reasoned that in the case of a null finding, such pathways might conceal a genuine causal relation between these variables—an example of *masking* (in the epidemiological sense), where the operation of one mechanism hides the effect of another (Illari 2011).

Illari's paper explains why I and a co-researcher (see Acknowledgments) found the proof-of-concept approach used by Trofimov and Williamson (2025) lacking as a definitive way to analyze the mask efficacy dataset. As Illari notes: "Without some reason to be confident that there are no confounders, evidence of a difference-making [that is, associative] relation between C and E does not establish a causal claim" (2011, 150).

Illari's attention to confounding resonated with the finding in this study that quantitative evidence of association (or its absence) could not meaningfully be combined with evidence of mechanism until a thorough, epidemiologically informed analysis of sources of bias had been undertaken. I developed the concept of *mechanisms of bias,* which considers, for each primary study, how and in what direction different biases might act—and hence, how they might mask or exaggerate a causal relationship. This work was fiendishly complex as a result of the sheer size and heterogeneity of the empirical dataset; it required the development of AI tools to identify and analyze multiple interacting biases in hundreds of primary studies (details will be reported separately).

In example 6 (Doung-Ngern et al. 2020) in Box 1, for example, an adjusted odds ratio of 0.23 for consistent mask use had to be read against the finding that those who always masked also kept their distance, spent less time with the index case, and washed their hands more often. Masking thus indexed a cluster of covarying protective behaviors, rather than an isolated exposure, and the stepwise, data-driven adjustment the authors used gave only partial assurance that the cluster had been disentangled.

4.2 Causal Complexity

Building on earlier work in this broader tradition, Illari and Russo (forthcoming) have recently argued that causal mechanisms in medicine are often complex and distributed, hence a comprehensive account of them is rarely achievable (but that a partial account is a lot better than nothing). Elsewhere, Russo et al. (2024) talk about the *causal mosaic*—a patchwork of interdependent causal pathways rather than a single, isolatable (C→E) cause.

This framing fitted the mask efficacy dataset well, since mask efficacy involves multilevel mechanisms from molecular to behavioral to societal and (sometimes) multiple parallel causal pathways (see, for example, the many variables that need to be taken into

account in example 6 in Box 1). Some of these mechanisms were easy to identify, others less so. I found the notion of "fallible heuristics" ("use what you know to suggest where to search next") helpful in guiding the search for more pieces of the mechanistic puzzle (Illari and Russo, forthcoming).

4.3 Intertwining

Whereas Evidential Pluralism separates associations and mechanisms, Illari and Russo argue for a return to evidential pluralism's traditional framing of these concepts as *intertwined:*

> In running any clinical trial or other correlational study, we need to choose variables, and how and when to measure them. … It is *not true* that you can get evidence of mechanism, see whether it is any good, and also evidence of correlation, and see whether it is any good; and then somehow put them together. (Illari and Russo, forthcoming; emphasis in original)

As Illari and Russo (forthcoming) show in relation to the contested topic of vitamin D deficiency in nonskeletal diseases, mechanistic assumptions necessarily inform the design of studies to test associations if those studies are to be done well. They offer a framework for considering, in relation to disease mechanisms and drugs intended to address them, (a) whether any mechanism is posited; (b) whether there is agreement about any posited mechanism(s); (c) whether multiple different mechanisms are posited; (d) whether variables measured are comparable across studies; and (e) how the measures might (or might not) be related to the mechanism posited (hence the extent to which they should be trusted).

The role of mechanistic reasoning in informing RCT design is well established (see, for example, the Medical Research Council guidance on design of complex interventions [Skivington et al. 2021]) but surfacing and fully examining such reasoning is not conventionally seen as mission-critical. The CONSORT reporting standards, for example, ask simply for an "explanation why the intervention might work" (Hopewell et al. 2025, supplementary materials, 8).

Illari and Russo's notion of intertwining provides a way in to systematically questioning the intervention design and choice of other variables in mask efficacy studies. Benjamin J. Cowling et al.'s (2009) household trial (example 1 in Box 1), for example, found no overall difference between arms, but a prespecified subgroup in which masking began within 36 hours of the index case developing symptoms showed a significant effect for masks combined with hand hygiene against control (adjusted odds ratio 0.33, 95% CI 0.13–0.87). That subgroup was not a post hoc trawl. It was specified in advance because viral shedding peaks in the first two days of symptoms (see example 2 in Box 1 [Suess et al. 2010]). In other words, a mechanistic hypothesis had been built into the design of a study that the framework would classify and appraise as a study of association. Appraising the association first and turning to mechanisms afterward would have discarded the only comparison in the trial capable of detecting an effect.

4.4 Biases and Confounders as Mechanisms

While traditional evidential pluralism proved more helpful in this analysis than Evidential Pluralism, it is worth being clear exactly what it contributed and what it did not. Evidential pluralism is a tradition rather than a method. The traditional approach offered orientation—a set of commitments about what causal inference in complex domains involves—and left the question of *how* to go about such inference entirely open. Applying this approach meant building something.

What we built came from noticing where the mechanistic evidence was actually doing its work. EBM+ set out to raise the status of mechanistic evidence about how an intervention produces its effect, and that is the evidence its frameworks teach us to look for. But in this dataset, the mechanistic evidence that mattered most was not about how masks work. It was about how the studies of masks went wrong: why an adjusted odds ratio of 0.23 could not be taken at face value (see the description of the Doung-Ngern et al. [2020] study in section 4.1), or why a trial reporting no effect had been incapable of detecting one (see the example of the hajj trial [Alfelali et al. 2020] in section 3.3).

Biases and confounders are not background noise to be adjusted away before the evidential work begins. They arise from causal processes; they have direction and magnitude; and they determine what a given study can and cannot detect. They are, in other words, mechanisms—and proper objects of the same appraisal as the mechanisms through which an intervention is supposed to act. Approaching them this way was what allowed the apparent inconsistencies in the mask literature to be resolved.

4.5 The Place of Expertise

Illari and Russo (forthcoming), following traditional evidential pluralism, treat expertise very differently. Their tools, developed from a chapter they coauthored in Parkkinen et al. (2018), are designed around the division of labor already operating in agencies such as the UK National Institute for Health and Care Excellence (NICE), where staff without domain expertise extract information from published studies and provide this to a panel of domain experts who assess it. Crucially, the tools do not score. Information is presented but not weighed, and the red flags are intended only to draw the panel's attention: It is for the experts to determine whether a flagged feature is a defect or a reasonable choice, given the prevailing state of the science. The rationale is epistemological rather than administrative. Whether mechanistic evidence is any good depends on what else is known, and this contextual quality is, as Illari and Russo note, particularly acute when one is interpreting an absence of evidence.

Expertise in this approach is thus *constitutive* of appraisal rather than a check upon it, and that is what our own experience required. The judgments of the methodologists we consulted were built into the AI tools we developed for the mask reviews, which were then iteratively refined as we analyzed the outputs alongside those experts. Thus, expert judgment shaped what the analysis looked for, rather than providing a downstream check on it once complete.

In sum, the traditional, broader version of evidential pluralism proved helpful for multiple reasons: (a) it required confounding to be addressed in parallel with mechanisms of action; (b) it allowed the causal complexity of the dataset to be examined; (c) it emphasized that

associations and mechanisms were empirically interconnected; (d) it enabled us to treat bias and confounding as mechanisms; and (e) it harnessed expertise as part of the process. However, its operationalization had to be worked out in a bespoke way to fit the dataset. The method we developed shares some core principles with that described by Illari and Russo (forthcoming) for combining empirical evidence of association and mechanism in vitamin D deficiency studies, but it differs in detail as it is rooted in the topic-specific challenges of a different empirical dataset.

5. Discussion

This paper has outlined the features of two approaches to evidential pluralism and described how they performed empirically on a dataset of mask efficacy studies. I believe it makes three contributions to the knowledge base on evidential pluralism.

First, I have described how Evidential Pluralism was useful conceptually but unhelpful empirically for a predominantly quantitative dataset characterized by causal complexity, heterogeneity, and extensive confounding. I have proposed explanations for the mismatch between concept and application. The framework is asymmetric, oriented toward explaining positive associations rather than interrogating null ones, and it offers no account of the inference that mattered most here—that mechanistic knowledge may establish that a study reporting a null finding was incapable of detecting the effect it tested for. It presupposes relatively simple and stable causal structures. It treats evidence of association and evidence of mechanism as separable streams, to be appraised independently and combined afterward. And it offers limited practical guidance for analyzing confounding and bias.

Second, I have mobilized the traditional, broader version of evidential pluralism and shown that concepts from its early literature, along with more recent extensions of them, did analytic work that Evidential Pluralism could not: Illari's (2011) insistence that identifying a mechanism strengthens but does not establish a causal claim, and that confounding must be disambiguated alongside it; the account of causal structures as complex, distributed and mosaic-like (Russo et al. 2024); Illari and Russo's (forthcoming) argument that mechanistic assumptions are already embedded in the design of studies testing association; and a constitutive perspective on expertise.

Third, I have argued that neither Evidential Pluralism nor traditional evidential pluralism supplies what a dataset of this kind most requires. Biases and confounders are not background noise to be adjusted away before the evidential work begins. They arise from causal processes with direction and magnitude, and they determine what a study is capable of detecting. They are therefore proper objects of the same mechanistic appraisal as the mechanisms through which an intervention is supposed to act. Approaching them this way was what allowed the apparent inconsistencies between mask efficacy studies to be resolved; the resulting method, mechanism-informed narrative synthesis (MINS), will be reported in detail separately.

The main strength of this analysis is that it emerged from a systematic and determined attempt to make evidential pluralism work for a uniquely complex and contested empirical dataset. The question of whether and in what circumstances (what kind of) masks are effective in controlling the spread of respiratory disease has been presented as one of the most controversial topics in public health, as demonstrated in the Covid-19 pandemic

(Muller 2021; Cash-Goldwasser et al. 2023) and more recently Andes hantavirus (Milton et al. 2026). At the heart of this controversy lie two philosophical questions: What role, if any, should evidence of mechanism play in assessing causality, and how should confounding and other sources of bias be handled in studies of association? Mask efficacy is thus a paradigmatic case of how to assess causality in epidemiological studies. To the extent that I have made philosophical progress with this dataset, my findings may apply more widely.

The main limitation is that I chose only two approaches to analyze in depth. I could have included many more philosophical approaches for combining different evidence types, including Cartwright's early work (1983) on the need for multiple, context-sensitive sources of evidence and her group's later work on scientific "tangles" (Cartwright et al. 2022); Caterina Marchionni's work on mechanisms in policy (Kuorikoski and Marchionni 2016; Marchionni and Reijula 2019); Michael Begun's explanatory-eliminative approach to amalgamation of complex and conflicting evidence (a philosophical PhD that uses airborne transmission of respiratory diseases and mask efficacy as one example) (Begun 2024); Jonathan Fuller's "argument framework" for synthesizing evidence on causality (again using mask efficacy as an example) (Fuller et al. 2024); and Rani Lill Anjum and the Cause Health group's work on patient-focused causal reasoning in healthcare (Anjum et al. 2020, 2025). I could also have included work on mechanisms by nonphilosophers, including Ray Pawson's work on realist review (Pawson 2024), David Ogilvie's group on methodological pluralism in public health reviews, using the "dry stone wall" metaphor (Ogilvie et al. 2020); the Cochrane-Campbell group's methodology for synthesizing intervention evidence from multiple sources (Calonge et al. 2023); and a novel approach from within EBM—the SPICE-GRADE approach for simultaneous processing of indirect causal evidence (Eachempati et al. 2026).

There is a further difficulty, which our dataset raised but did not resolve. Evidence-based medicine has built an elaborate apparatus for appraising evidence of association, including risk-of-bias tools, GRADE, reporting standards, and decades of methodological work on what makes a trial trustworthy. Nothing comparable exists for evidence of mechanism. Illari's (2011) warning that identifying a mechanism strengthens a causal claim without establishing it, and Howick's (2011) that mechanistic reasoning is prone to just-so stories, both point at the problem: Mechanistic accounts vary enormously in how well evidenced they are, and we lack agreed adjudication methods.

Our dataset illustrates the difficulty in both directions. The mechanisms that proved decisive—airborne transmission, nonadherence, leakage at the face seal, the timing of viral shedding—were supported by experimental work of a quality that would satisfy most readers. But mechanistic accounts of comparable superficial plausibility were available on either side of the mask debate throughout the pandemic and were deployed with confidence by people who disagreed profoundly. Nothing in either approach told us how to distinguish them. A framework that requires evidence of mechanism to be brought in, while offering no systematic way of appraising what is brought, does not merely fail to help. It risks conferring the authority of a method on whatever mechanistic account happens to be to hand.

It is worth noting that the empirical work of synthesizing evidence on mask efficacy was resource intensive and required an AI-supported method we had to build ourselves. This may seem paradoxical. But the complexity was in the evidence, not in the method we imposed on it: some 400 primary studies, each subject to multiple biases operating in different directions and to different degrees. Any approach adequate to such a dataset must

be commensurate with it. The question is not whether a method is demanding but whether what it demands is what the evidence requires—and the effort Evidential Pluralism called for went into sorting studies into channels rather than into establishing what each study was capable of showing. Clinicians and bench scientists integrate mechanistic and associative reasoning every day and mostly do it well—considering what a trial showed, what is known about how the drug behaves, whether this patient resembles those enrolled—without any framework at all, and without finding it especially difficult. Formalizing that reasoning made it considerably harder—but that added difficulty is the point, since intuitive integration, however fluent, cannot catch its own errors. Catching those errors requires what is missing in everyday reasoning: a means of appraising mechanistic evidence, and an account of biases as mechanisms.

6. Conclusion

Mechanisms are increasingly recognized as essential components of the knowledge base in medicine and healthcare. The EBM+ movement has helped shift medical discourse beyond EBM's original mechanism-blind hierarchy of evidence (Stegenga 2014; Wilde 2023). But as this study illustrates, the time to go beyond proof of concept is overdue.

Defenders of Evidential Pluralism may reasonably object that the framework can accommodate everything I have said it cannot handle. Confounding can be represented within it. So can attenuation, and so can null findings. I accept all this. But the objection and the reply are answers to different questions. A formalism is *expressively adequate* if its vocabulary is rich enough to represent the distinctions that matter in a domain. It is *methodologically determinate* if, faced with a large and heterogeneous body of studies, it tells me what to do next and what to conclude. The first is a property of a notation and is not hard to come by. The second is a property of a method and is very hard indeed. Formalization risks delivering the first while looking as though it has delivered the second—because a notation that can express a distinction looks, on the page, exactly like a procedure for dealing with it. This, I think, is why proof of concept has proved so durable in this literature. Where the causal conclusion is already known, expressive adequacy is all you need: The formalism restates an argument that was reached by other means and appears thereby to have produced it.

Despite extensive searching, I have found no published, successful empirical application of Evidential Pluralism. However, I am not arguing that none is possible, and I know that Williamson is working on further examples. What would count as a successful application would involve a systematically assembled dataset, where the answer was not known at the outset,[3] in which the steps were followed in sequence, and the judgments reached without recourse to resources the framework does not supply. The framework's proponents are better placed than I am to produce such an application, and this paper should be read as an invitation to do so.

[3] A case on which everyone is already agreed will not be able to discriminate between different approaches to combining evidence, since any framework will arrive at an agreed conclusion. That the mask evidence was contested is what made it informative.

Acknowledgments
All views expressed in this paper are mine alone. I am grateful to the following people for the many conversations and insights that have influenced the writing of this paper: Jon Williamson, Rebecca Helm, Luana Poliseli, and Sahanika Ratnayake were all employed on the Interdisciplinary Systematic Review: Mechanistic Evidence and Epistemic Justice project and contributed to searching, appraisal, and discussions about the mask efficacy papers and how to analyze them. In particular, Sahanika Ratnayake was second reviewer on the dataset of RCTs and observational studies of mask efficacy. Alexandra Trofimov, a member of Jon Williamson's wider team, also attended some meetings and contributed to discussions. Federica Russo was a member of the external advisory group for that project and Nancy Cartwright was a presenter at, and regular attender of, the monthly seminars funded by the grant. Phyllis Illari was an invited speaker at a two-day workshop organized by the project in March 2026, where she presented some of the ideas on which I touch briefly in this paper. Nancy Cartwright, Cheryl Misak, and I held many conversations about philosophy of evidence. I thank Cheryl Misak, Nancy Cartwright, John Worrall, and an early career researcher for helpful comments on previous drafts of this paper.

Disclosure Statement
I am a co-investigator with Jon Williamson and Rebecca Helm on the UKRI Interdisciplinary Systematic Review: Mechanistic Evidence and Epistemic Justice project. As described in the paper, the tensions I describe in section 3 are ones that I lived and breathed while trying to deliver on that research study. See also Acknowledgments.

References
Abdin, Ahmad Y., Daniel Auker-Howlett, Jürgen Landes, Glorjen Mulla, Claus Jacob, and Barbara Osimani. 2019. "Reviewing the Mechanistic Evidence Assessors E-Synthesis and EBM+: A Case Study of Amoxicillin and Drug Reaction with Eosinophilia and Systemic Symptoms (DRESS)." *Current Pharmaceutical Design* 25, no. 16: 1866–1880. https://doi.org/10.2174/1381612825666190628160603.

Alfelali, Mohammad, Elizabeth A. Haworth, Osamah Barasheed, et al. 2020. "Facemask Against Viral Respiratory Infections Among Hajj Pilgrims: A Challenging Cluster-Randomized Trial." *PLoS One* 15, no. 10: e0240287. https://doi.org/10.1371/journal.pone.0240287.

Anjum, Rani Lill, Samantha Copeland, and Elena Rocca. 2020. *Rethinking Causality, Complexity and Evidence for the Unique Patient: A Cause Health Resource for Healthcare Professionals and the Clinical Encounter*. Springer.

Anjum, Rani Lill, Christine Price, and Elena Rocca. 2025. "When Patient Voices Get Lost in Evidence Hierarchies: A Testimony of Rare Adverse Events and Participatory Epistemic Injustice in Drug Safety Monitoring." *Social Epistemology* 39, no. 2: 187–201. https://doi.org/10.1080/02691728.2024.2400079.

Aronson, Jeffrey K., Adam La Caze, Michael P. Kelly, Veli-Pekka Parkkinen, and Jon Williamson. 2018. "The Use of Mechanistic Evidence in Drug Approval." *Journal of Evaluation in Clinical Practice* 24, no. 5: 1166–1176. https://doi.org/10.1111/jep.12960.

Auker-Howlett, Daniel, and Jon Williamson. 2021. "Vaccination Uptake Interventions: An EBM+ Approach." *Argumenta* 7, no. 1: 79–96. https://doi.org/10.14275/2465-2334/202113.AUK.

Begun, Michael. 2024. "Epistemic Issues in Evidence Amalgamation." PhD diss., University of Pittsburgh.

Borgerson, Kirstin. 2009. "Valuing Evidence: Bias and the Evidence Hierarchy of Evidence-Based Medicine." *Perspectives in Biology and Medicine* 52, no. 2: 218–233. https://doi.org/10.1353/pbm.0.0086.

Bradford Hill, Austin. 1965. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58, no. 5: 295–300. https://doi.org/10.1177/003591576505800503.

Calonge, Ned, Paul G. Shekelle, Douglas K. Owens, et al. 2023. "A Framework for Synthesizing Intervention Evidence from Multiple Sources into a Single Certainty of Evidence Rating: Methodological Developments from a US National Academies of Sciences, Engineering, and Medicine Committee." *Research Synthesis Methods* 14, no. 1: 36–51. https://doi.org/10.1002/jrsm.1582.

Campbell, Donald T. 1987. "Evolutionary Epistemology." In *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge,* edited by Gerard Radnitzky and W.W. Bartley III. Open Court.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford University Press.

———. 2021. "What Counts as Evidence for a Mechanism? And Why?" *The Reasoner* 15, no. 6. https://riviste.unimi.it/index.php/thereasoner/article/view/24261.

Cartwright, Nancy, Jeremy Hardie, Eleonora Montuschi, Matthew Soleiman, and Ann C. Thresher. 2022. *The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity*. Oxford University Press.

Cash-Goldwasser, Shama, Arthur L. Reingold, Stephen P. Luby, Lisa A. Jackson, and Thomas R. Frieden. 2023. "Masks During Pandemics Caused by Respiratory Pathogens—Evidence and Implications for Action." *JAMA Network Open* 6, no. 10: e2339443. https://doi.org/10.1001/jamanetworkopen.2023.39443.

Clarke, Brendan, Donald Gillies, Phyllis Illari, Federica Russo, and Jon Williamson. 2013. "The Evidence That Evidence-Based Medicine Omits." *Preventive Medicine* 57, no. 6: 745–747. https://doi.org/10.1016/j.ypmed.2012.10.020.

———. 2014. "Mechanisms and the Evidence Hierarchy." *Topoi* 33, no. 2: 339–360. https://doi.org/10.1007/s11245-013-9220-9.

Cowling, Benjamin J., Kwok-Hung Chan, Vicky J. Fang, et al. 2009. "Facemasks and Hand Hygiene to Prevent Influenza Transmission in Households: A Cluster Randomized Trial." *Annals of Internal Medicine* 151, no. 7: 437–446. https://doi.org/10.7326/0003-4819-151-7-200910060-00142.

Doung-Ngern, Pawinee, Rapeepong Suphanchaimat, Apinya Panjangampatthana, et al. 2020. "Case-Control Study of Use of Personal Protective Measures and Risk for SARS-CoV 2 Infection, Thailand." *Emerging Infectious Diseases* 26, no. 1): 2607–2616. https://doi.org/10.3201/eid2611.203003.

Dowe, Phil. 2000. *Physical Causation*. Cambridge University Press.

Eachempati, Prashanti, Gordon Guyatta, Per Olav Vandvika, et al. 2026. "SPICE-GRADE: Simultaneous Processing of Indirect Causal Evidence in Complex Pathways Using GRADE—An Exploratory Case Study." *Journal of Clinical Epidemiology* 194, 112219. https://doi.org/10.1016/j.jclinepi.2026.112219.

Frank, Christina, Mirko Faber, and Klaus Stark. 2016. "Causal or Not: Applying the Bradford Hill Aspects of Evidence to the Association Between Zika Virus and Microcephaly." *EMBO Molecular Medicine* 8, no. 4: 305–307. https://doi.org/10.15252/emmm.201506058.

Fuller, Jonathan, Benjamin Chin-Yee, and Ross E.G. Upshur. 2024. "The Argument Framework Is a Flexible Approach to Evidence in Healthcare." *Nature Medicine* 30: 2113–2116. https://doi.org/10.1038/s41591-024-02930-x.

Gillies, Donald. 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.

Goldenberg, Maya J. 2009. "Iconoclast or Creed? Objectivism, Pragmatism, and the Hierarchy of Evidence." *Perspectives in Biology and Medicine* 52, no. 2: 168–187. https://doi.org/10.1353/pbm.0.0080.

Greenhalgh, Trisha, David Fisman, Danielle J. Cane, Matthew Oliver, and Chandini Raina Macintyre. 2022. "Adapt or Die: How the Pandemic Made the Shift from EBM to EBM+ More Urgent." *BMJ Evidence-Based Medicine* 27, no. 5: 253–260. https://doi.org/10.1136/bmjebm-2022-111952.

Greenhalgh, Trisha, Rebecca Helm, Luana Poliseli, et al. 2025. "Protocol: A Systematic Review+ (SR+) to Combine Associative and Mechanistic Evidence on the Efficacy of Face Masks in Reducing Transmission of Respiratory Diseases." *Systematic Reviews* 14, no. 1, 226. https://doi.org/10.1186/s13643-025-02973-2.

Greenhalgh, Trisha, Sahanika Ratnayake, Rebecca Helm, Luana Poliseli, and Jon Williamson. 2026a. "The Randomised Controlled Trial as Cultural Product: An Illustrative Review Using Mask Efficacy Trials." *SSM – Qualitative Research in Health* 9, 100772. https://doi.org/10.1016/j.ssmqr.2026.100772.

———. 2026b. "Synthesis Challenges in Observational Studies of Face Mask Efficacy: A Critical Analysis of Published Systematic Reviews." *Research Synthesis Methods* 17, no. 4: 714–733. https://doi.org/10.1017/rsm.2026.10072.

Guyatt, Gordon, John Cairns, David Churchill, et al. 1992. "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine." *JAMA* 268, no. 17: 2420–2425. https://doi.org/10.1001/jama.1992.03490170092032.

Han, Paul K.J. 2021. *Uncertainty in Medicine: A Framework for Tolerance*. Oxford University Press.

Harding, Sandra, and Merrill B Hintikka. 2003. *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. Vol. 161. Springer Science & Business Media.

Helm, Rebecca K., Trisha Greenhalgh, Jon Williamson, and Alexandra Trofimov. 2026. "When Trial Interventions Are Likely to Reduce the Impact of Rape Myths on Juror Decisions: A Mechanism-Informed Evidence Synthesis." *Journal of Criminal Justice* 104, 102634. https://doi.org/10.1016/j.jcrimjus.2026.102634.

Holman, Bennett. 2019. "Philosophers on Drugs." *Synthese* 196, no. 11: 4363–4390. https://doi.org/10.1007/s11229-017-1642-2.

Hopewell, Sally, An-Wen Chan, Gary S. Collins, et al. 2025. "CONSORT 2025 Statement: Updated Guideline for Reporting Randomised Trials." *The Lancet* 405, no. 10489: 1633–1640. https://doi.org/10.1016/S0140-6736(25)00672-5.

Howick, Jeremy. 2011. "Exposing the Vanities—and a Qualified Defense—of Mechanistic Reasoning in Health Care Decision Making." *Philosophy of Science* 78, no. 5: 926–940. https://doi.org/10.1086/662561.

Illari, Phyllis McKay. 2011. "Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis." *International Studies in the Philosophy of Science* 25, no. 2: 139–157. https://doi.org/10.1080/02698595.2011.574856.

Illari, Phyllis McKay, and Federica Russo. Forthcoming. "Practical Tools for Using Evidence of Mechanisms to Establish Correlations." *Journal of Evaluation in Clinical Practice*.

Krauer, Fabienne, Maurane Riesen, Ludovic Reveiz, et al; WHO Zika Causality Working Group. 2017. "Zika Virus Infection as a Cause of Congenital Brain Abnormalities and Guillain–Barré Syndrome: Systematic Review." *PLoS Medicine* 14, no. 1: 1–27. https://doi.org/10.1371/journal.pmed.1002203.

Kuorikoski, Jaakko, and Caterina Marchionni. 2016. "Evidential Diversity and the Triangulation of Phenomena." *Philosophy of Science* 83, no. 2: 227–247. https://doi.org/10.1086/684960.

Kutter, Jasmin S., Dennis de Meulder, Theo M. Bestebroer, et al. 2021. "SARS-CoV and SARS-CoV-2 Are Transmitted Through the Air Between Ferrets over More Than One Meter Distance." *Nature Communications* 12, no. 1, 1653. https://doi.org/10.1038/s41467-021-21918-6.

Lee, Seungjae, Tark Kim, Eunjung Lee, et al. 2020. "Clinical Course and Molecular Viral Shedding Among Asymptomatic and Symptomatic Patients with SARS-CoV-2 Infection in a Community Treatment Center in the Republic of Korea." *JAMA Internal Medicine* 180, no. 11: 1447–1452. https://doi.org/10.1001/jamainternmed.2020.3862.

Marchionni, Caterina, and Samuli Reijula. 2019. "What Is Mechanistic Evidence, and Why Do We Need It for Evidence-Based Policy?" *Studies in History and Philosophy of Science Part A* 73: 54–63. https://doi.org/10.1016/j.shpsa.2018.08.003.

Milton, Donald K., Trisha Greenhalgh, David N. Fisman, Amanda Kvalsvig, Lidia Morawska, and Jonathan M. Samet. 2026. "Hantavirus Outbreak Should Reset WHO's Default Approach to Airborne Risk." *BMJ* 393, s919. https://doi.org/10.1136/bmj.s919.

Muller, Seán M. 2021. "Masks, Mechanisms and Covid-19: The Limitations of Randomized Trials in Pandemic Policymaking." *History and Philosophy of the Life Sciences* 43, no. 2: 1–5. https://doi.org/10.1007/s40656-021-00403-9.

Ogilvie, David, Adrian Bauman, Louise Foley, Cornelia Guell, David Humphreys, and Jenna Panter. 2020. "Making Sense of the Evidence in Population Health Intervention Research: Building a Dry Stone Wall." *BMJ Global Health* 5, no. 12, e004017. https://doi.org/10.1136/bmjgh-2020-004017.

Parkkinen, Veli-Pekka, Christian Wallmann, Michael Wilde, et al. 2018. *Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures*. Springer Nature.

Pawson, Ray. 2024. *How to Think Like a Realist: A Methodology for Social Science*. Edward Elgar Publishing.

Pfaff, Holger, and Jochen Schmitt. 2023. "Reducing Uncertainty in Evidence-Based Health Policy by Integrating Empirical and Theoretical Evidence: An EbM+Theory Approach." *Journal of Evaluation in Clinical Practice* 29, no. 8: 1279–1293. https://doi.org/10.1111/jep.13890.

Rasmussen, Sonja A, Denise J. Jamieson, Margaret A. Honein, and Lyle R. Petersen. 2016. "Zika Virus and Birth Defects—Reviewing the Evidence for Causality." *New England Journal of Medicine* 374, no. 20: 1981–1987. https://doi.org/10.1056/NEJMsr1604338.

Riboli, Elio, Frederick A. Beland, Dirk W. Lachenmeier, et al. 2023. "Carcinogenicity of Aspartame, Methyleugenol, and Isoeugenol." *The Lancet Oncology* 24, no. 8: 848–850. https://doi.org/10.1016/s1470-2045(23)00341-8.

Russo, Federica. 2006. "The Rationale of Variation in Methodological and Evidential Pluralism." *Philosophica* 77: 97–124. https://russofederica.wordpress.com/wp-content/uploads/2009/11/volume77-volledig.pdf.

Russo, Federica, Alex Broadbent, Brian Castellani, et al. 2024. "A Pluralistic (Mosaic) Approach to Causality in Health Complexity." In *The Routledge Handbook of Causality and Causal Methods,* edited by Phyllis Illari and Federica Russo. Routledge.

Russo, Federica, and Jon Williamson. 2007. "Interpreting Causality in the Health Sciences." *International Studies in the Philosophy of Science* 21, no. 2: 157–170. https://doi.org/10.1080/02698590701498084.

Sackett, David L., R. Brian Haynes, Gordon H. Guyatt, and Peter Tugwell. 1991. *Clinical Epidemiology: A Basic Science for Clinical Medicine,* 2nd ed. Little, Brown.

Salmon, Wesley C. 1998. *Causality and Explanation*. Oxford University Press.

Schmitt, Jean, and Jing Wang. 2022. "A Critical Review on the Role of Leakages in the Facemask Protection Against SARS-CoV-2 Infection with Consideration of Vaccination and Virus Variants." *Indoor Air* 32, no. 10, e13127. https://doi.org/10.1111/ina.13127.

Shan, Yafeng, and Jon Williamson. 2021. "Applying Evidential Pluralism to the Social Sciences." *European Journal for Philosophy of Science* 11: 1–27. https://doi.org/10.1007/s13194-021-00415-z.

Shepard, Thomas H. 1994. "'Proof' of Human Teratogenicity." *Teratology* 50, no. 2: 97–98. https://doi.org/10.1002/tera.1420500202.

Skivington, Kathryn, Lynsay Matthews, Sharon Anne Simpson, et al. 2021. "A New Framework for Developing and Evaluating Complex Interventions: Update of Medical Research Council Guidance." BMJ 374, n2061. https://doi.org/10.1136/bmj.n2061.

Smith, Martyn T., Kathryn Z. Guyton, Catherine F. Gibbons, et al. 2016. "Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis." *Environmental Health Perspectives* 124, no. 6: 713–721. https://doi.org/10.1289/ehp.1509912.

Stegenga, Jacob. 2014. "Down with the Hierarchies." *Topoi* 33, no. 2: 313–322. https://doi.org/10.1007/s11245-013-9189-4.

Suess, Thorsten, Udo Buchholz, Susann Dupke, et al. 2010. "Shedding and Transmission of Novel Influenza Virus A/H1N1 Infection in Households—Germany, 2009." *American Journal of Epidemiology* 171, no. 11: 1157–1164. https://doi.org/10.1093/aje/kwq071.

Trofimov, Alexandra, and Jon Williamson. 2025. "Applying Evidential Pluralism to Evidence-Based Law: EBL+." *Jurisprudence* 16, no. 4: 647–690. https://doi.org/10.1080/20403313.2025.2449758.

Wilde, Michael. 2023. "The EBM+ Movement." *International Journal of Biostatistics* 19, no. 2: 283–293. https://doi.org/10.1515/ijb-2022-0126.

Williamson, Jon. 2019. "Evidential Proximity, Independence, and the Evaluation of Carcinogenicity." *Journal of Evaluation in Clinical Practice* 25, no. 6: 955–961. https://doi.org/10.1111/jep.13226.

———. 2021a. "Establishing the Teratogenicity of Zika and Evaluating Causal Criteria." *Synthese* 198, no. 10: 2505–2518. https://doi.org/10.1007/s11229-018-1866-9.

———. 2021b. "The Feasibility and Malleability of EBM+." *Theoria* 36, no. 2: 191–209. https://doi.org/10.1387/theoria.21244.