Causal inference is not statistical inference

Jon Williamson
Philosophy Department and Centre for Reasoning
University of Kent

8 December 2022
The statistics wars and their casualties

1 Introduction

Causal enquiry. Establishing that A is a cause of B, or assessing whether A is a cause of B.

Key techniques for causal enquiry are couched as statistical. E.g.,

- Analysis of RCTs
- Analyses of cohort studies and other observational studies
- Meta-analysis
- Learning a structural equation model
- Learning a graphical causal model.

Two theses:

Weak. Statistical techniques are useful for causal enquiry.

Strong. Causal enquiry is a purely statistical problem.

There seems to be a trend from the weak thesis to the strong.

R.A. Fisher in *Statistical methods for research workers*:

> from the modern point of view, the study of the causes of variation of any variable phenomenon, from the yield of wheat to the intellect of man, should be begun by the examination and measurement of the variation which presents itself. (Fisher, 1925, p. 3.)

This is compatible with the weak thesis.

Fisher championed the use of both randomised (Fisher, 1935) and observational studies for causal enquiry.

Austin Bradford Hill can be viewed as another advocate of the weak thesis.

> Over a fairly wide expanse of clinical medicine in Great Britain the statistical approach has been accepted as useful (Hill, 1952, p. 113).

> The statistically guided therapeutic trial is not the only means of investigation and experiment, nor indeed is it invariably the best way of advancing knowledge of therapeutics. I commend it to you as *one* way, and, I believe, a useful way (Hill, 1952, p. 119).

Hill (1965) stressed the gap between association and causation.

> His 9 ‘viewpoints’ or indicators of causation help to bridge the gap.
>
> Several of these indicators are not statistical indicators.

Gillies (2019, Chapter 9) argues that Hill’s streptomycin trials, which began in 1946 and were perhaps the earliest published examples of medical RCTs, were far from a purely statistical approach to causal enquiry.

These days, however, the strong thesis predominates

The hierarchies of EBM and EBP focus almost exclusively on the analysis of RCTs.

Quantitative statistical methods now dominate causal enquiry in the social sciences.

The graphical causal modelling approach of Pearl and others conceives of causal relationships as inferrable statistical models with Markovian properties (Lauritzen, 1996).

> When attempting to make sense of data, statisticians are invariably motivated by causal questions. . . . The peculiar nature of these questions is that they cannot be answered, or even articulated, in the traditional language of statistics. In fact, only recently has science acquired a mathematical language we can use to express such questions, with accompanying tools to allow us to answer them from data.
>
> The development of these tools has spawned a revolution in the way causality is treated in statistics and in many of its satellite disciplines. (Pearl et al., 2016, p. xi.)

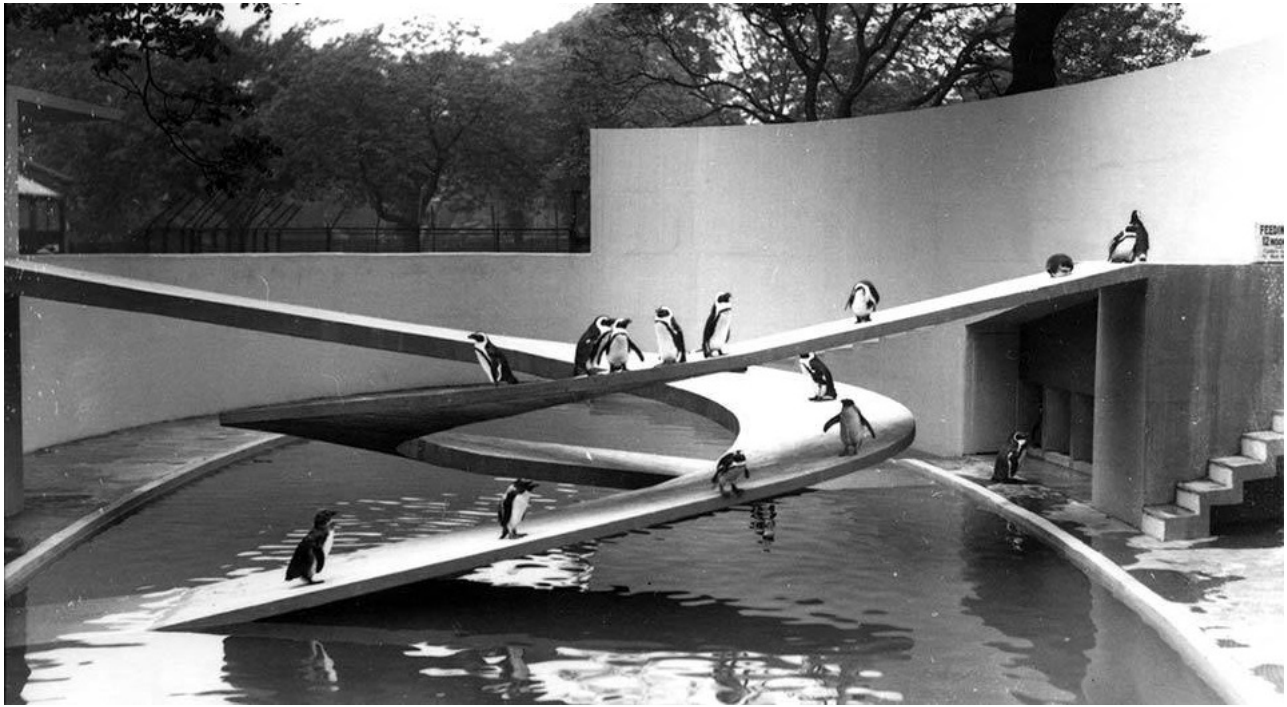
I.e., recent improvements in statistical methods motivate the strong thesis.

There's also a trend towards conflating causation with correlation conditional on potential confounders.

EG RCTs are touted as providing unbiased estimates of 'causal effects'.

Contents

2 Evidential Pluralism



Correlation is not Causation. A correlation might be explained by:

Causation	A is a cause of B.
Reverse causation	B is a cause of A.
Confounding (selection bias)	There is some confounder C that has not been adequately controlled for by the study.
Performance bias	Those in the A-group are identified and treated differently to those in the $\neg A$-group.
Detection bias	B is measured differently in the A-group in comparison to the $\neg A$-group.
Chance	Sheer coincidence, attributable to too small a sample.
Fishing	Measuring so many outcomes that there is likely to be a chance correlation between A and some such B.
Temporal trends	A and B both increase over time for independent reasons. E.g., prevalence of coeliac disease & spread of HIV.
Semantic relationships	Overlapping meaning. E.g., phthiasis, consumption, scrofula (all of which are TB).
Constitutive relationships	One variable is a part or component of the other.
Logical relationships	Measurable variables A and B are logically complex and logically overlapping. E.g., A is $C \wedge D$ and B is $D \vee E$.
Nomological relationships	E.g., conservation of total energy can induce a correlation between two energy measurements.
Mathematical relationships	E.g., mean and variance variables from the same distribution can be correlated.

If *A* is a cause of *B*, then there is some complex of mechanisms that:

Explains instances of *B* by invoking instances of *A*, and

Can account for the magnitude of the observed correlation.

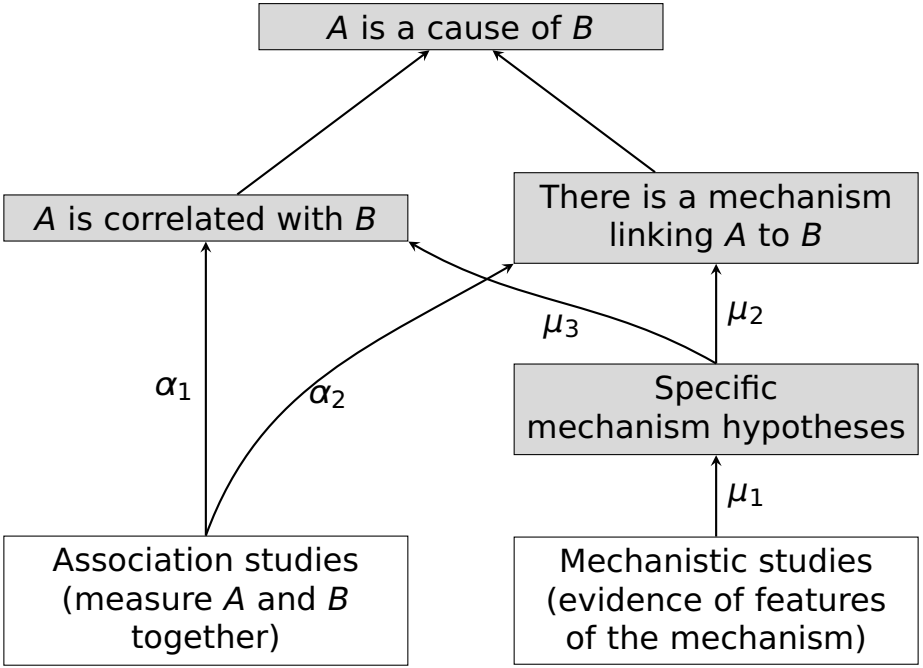
∴ In order to establish causation one needs to establish both:

The existence of an appropriate correlation.

The existence of an appropriate mechanism that can explain that correlation.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

This observation motivates **Evidential Pluralism**, a theory of causal enquiry:



Evidential Pluralism = object pluralism + study pluralism

Association and Mechanistic studies reinforce one another

Association and mechanistic studies have complementary strengths:

Association studies can be unreliable indicators of causality because of biases etc.

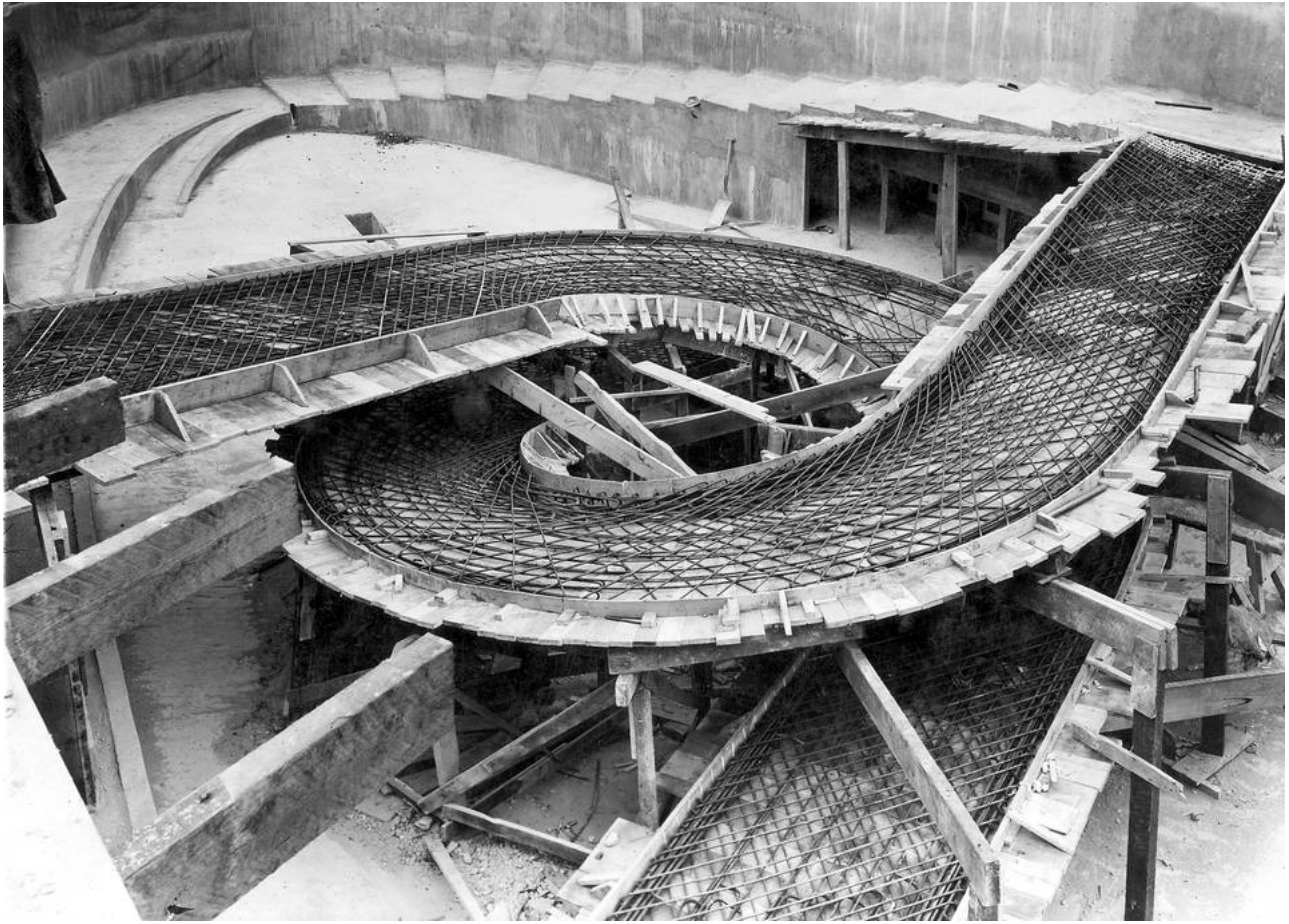
Mechanistic studies help to isolate the correlations that are genuinely causal.

Mechanistic studies can suffer from the problems of masking and complexity.

Association studies help to isolate the mechanisms that make a net difference.

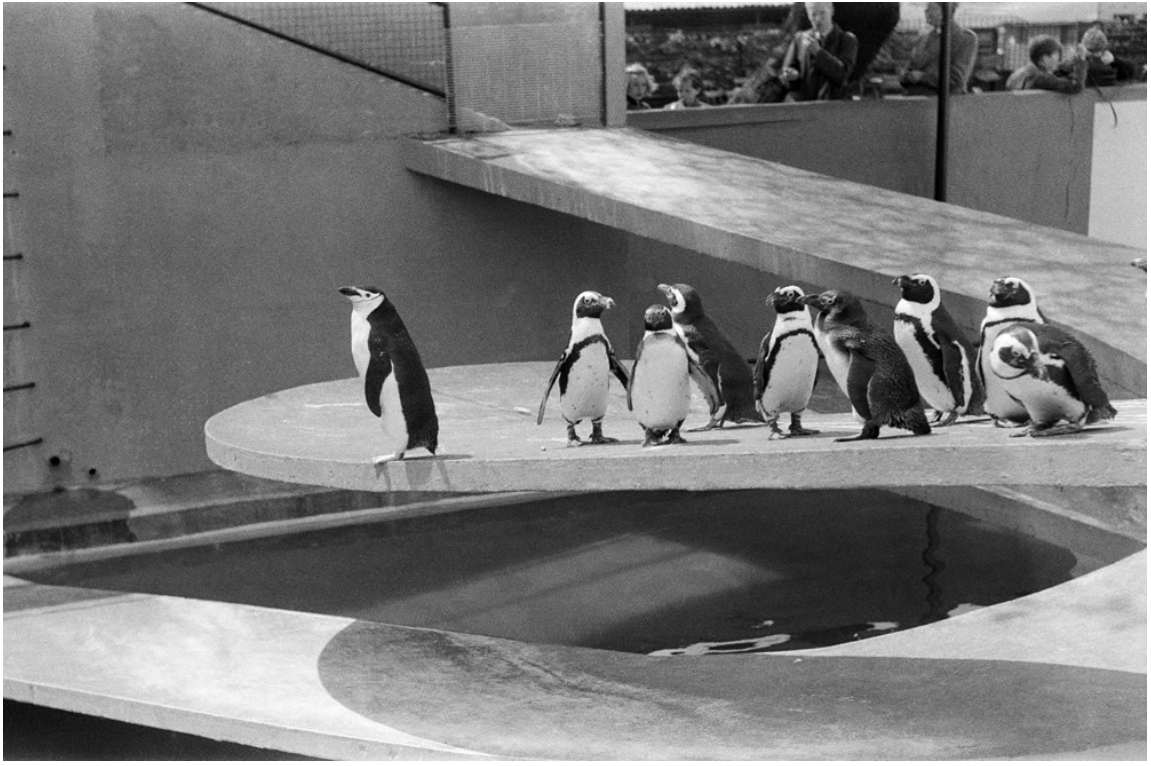
∴ Association studies and mechanistic studies reinforce each other.

Their combined evidential value is more than the sum of the parts.



Berthold Lubetkin’s 1934 London Zoo Penguin Pool, an early example of reinforced concrete.

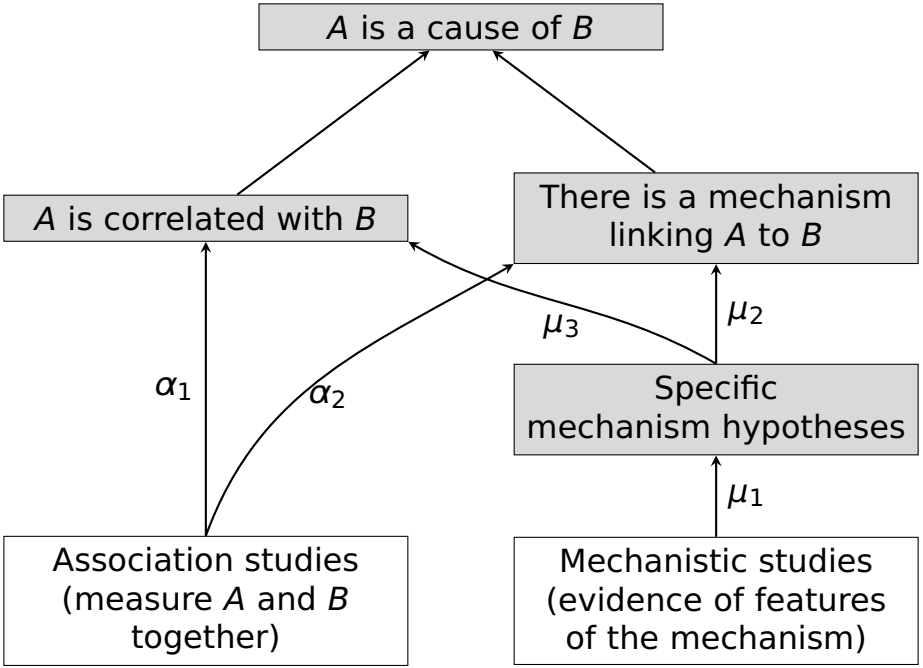
3 Evidential Pluralism in medicine



Present-day EBM focusses on the α-channels and takes mechanistic studies to be strictly inferior to association (clinical/epidemiological) studies.

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

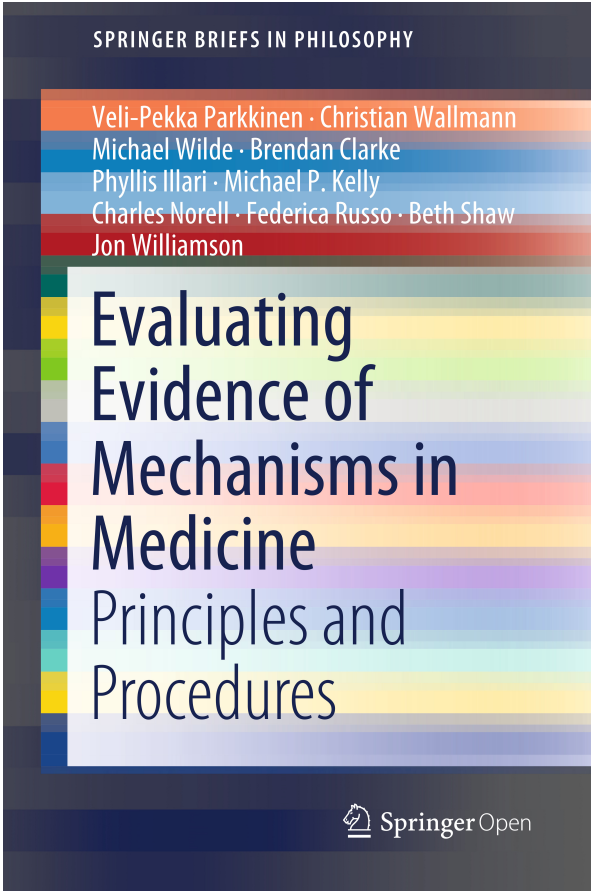
Evidential Pluralism takes the μ-channels to be important too.



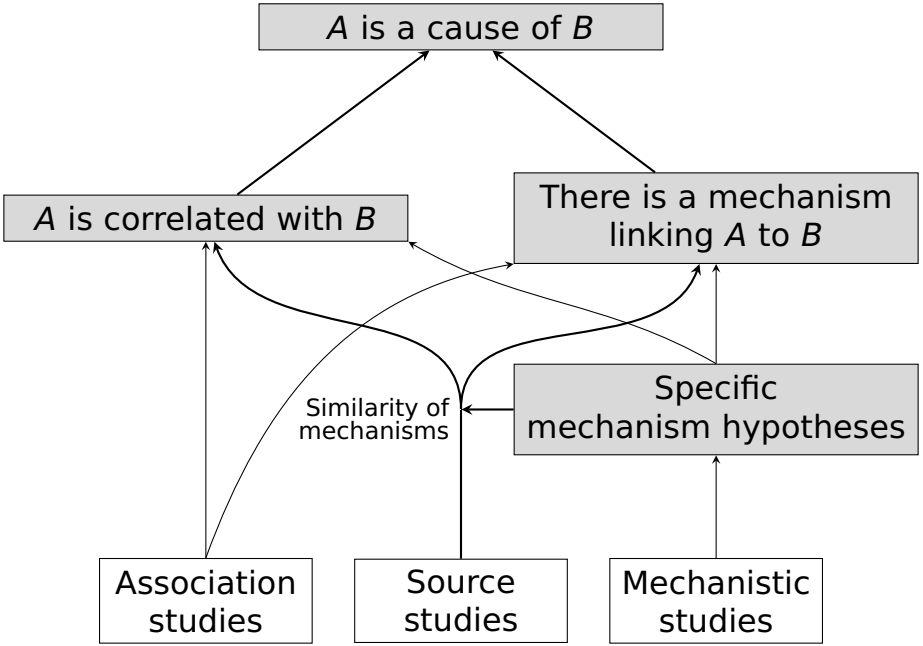
In medicine, this leads to an approach called EBM+.

EBM+ systematically scrutinises mechanistic studies alongside association studies.

EBM+ handbook (open access):

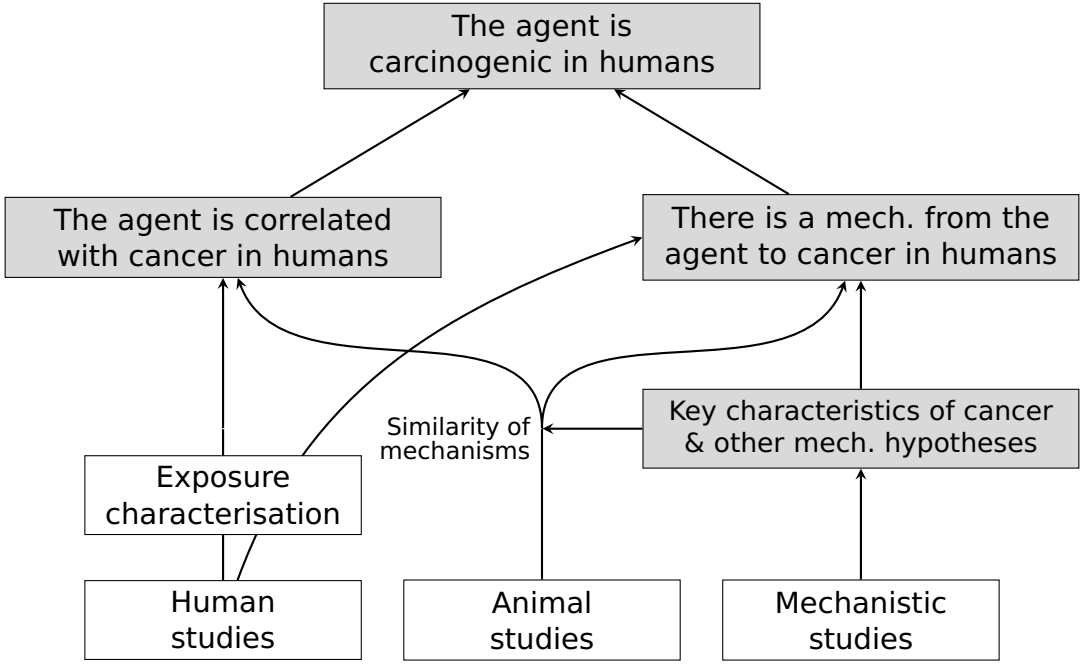


EP also provides an account of external validity:



Mechanistic evidence is routinely assessed in this way in exposure assessment:

EG IARC integrates human, animal and mechanistic studies.



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

It has been argued that EBM+ would have served us better than EBM during the pandemic.

EBM analysis

Adapt or die: how the pandemic made the shift from EBM to EBM+ more urgent

Trisha Greenhalgh ,[1] David Fisman,[2] Danielle J Cane,[3] Matthew Oliver ,[4] Chandini Raina Macintyre[5]

10.1136/bmjebm-2022-111952

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[3]Coalition for Healthcare Acquired Infection Reduction, Cambridge, Ontario, Canada
[4]Association of Professional Engineers and Geoscientists, Edmonton, Alberta, Canada
[5]The Biosecurity Program, Kirby Institute, University of New South Wales, Sydney, New South Wales, Australia

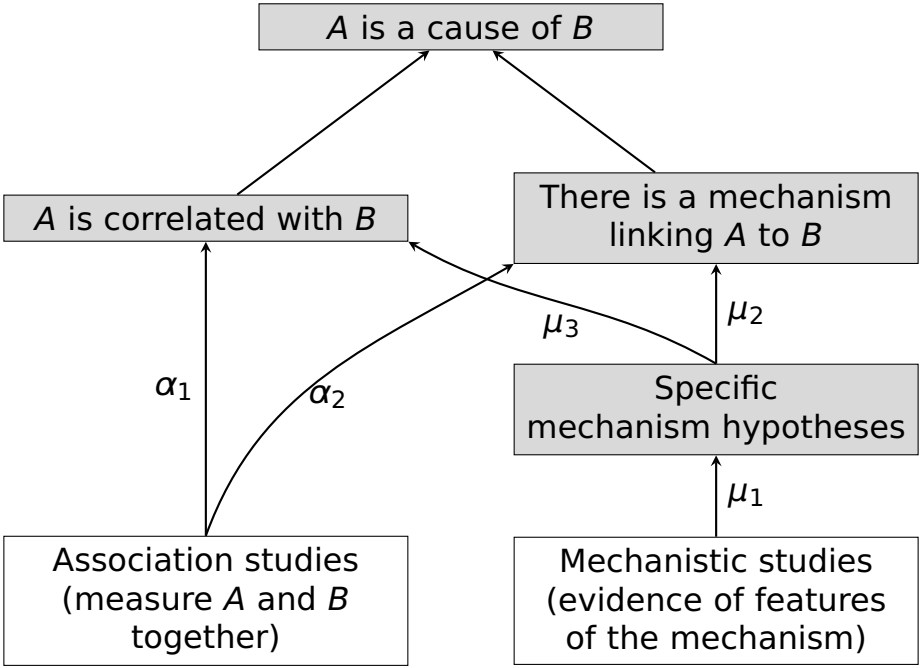
Abstract

Evidence-based medicine (EBM's) traditional methods, especially randomised controlled trials (RCTs) and meta-analyses, along with risk-of-bias tools and checklists, have contributed significantly to the science of COVID-19. But these methods and tools were designed primarily to answer simple, focused questions in a stable context where yesterday's research can be mapped more or less unproblematically onto today's clinical and policy questions. They have significant limitations when extended to complex questions about a novel pathogen causing chaos across multiple sectors in a fast-changing global context. Non-pharmaceutical interventions which combine material artefacts, human behaviour, organisational directives, occupational health and safety, and the built environment are a case in point: EBM's experimental, intervention-and—in some—prolonged sequelae. Effective and safe vaccines were produced rapidly, but uptake has been patchy and highly transmissible variants continue to spread and mutate. Coordinated disinformation campaigns have weakened the public health response.

Despite a quarter of a million scientific papers on COVID-19, some basic issues remain contested. How exactly does the virus spread? How effective are non-pharmaceutical interventions—masks, distancing, closure of buildings, remote working and learning, lockdowns—in reducing transmission, and what are their trade-offs? How can we make schools, hospitals and other public buildings safe? How can we protect workers and the public without closing down the economy? How cn we reduce the shocking inequalities that have characterised this pandemic?

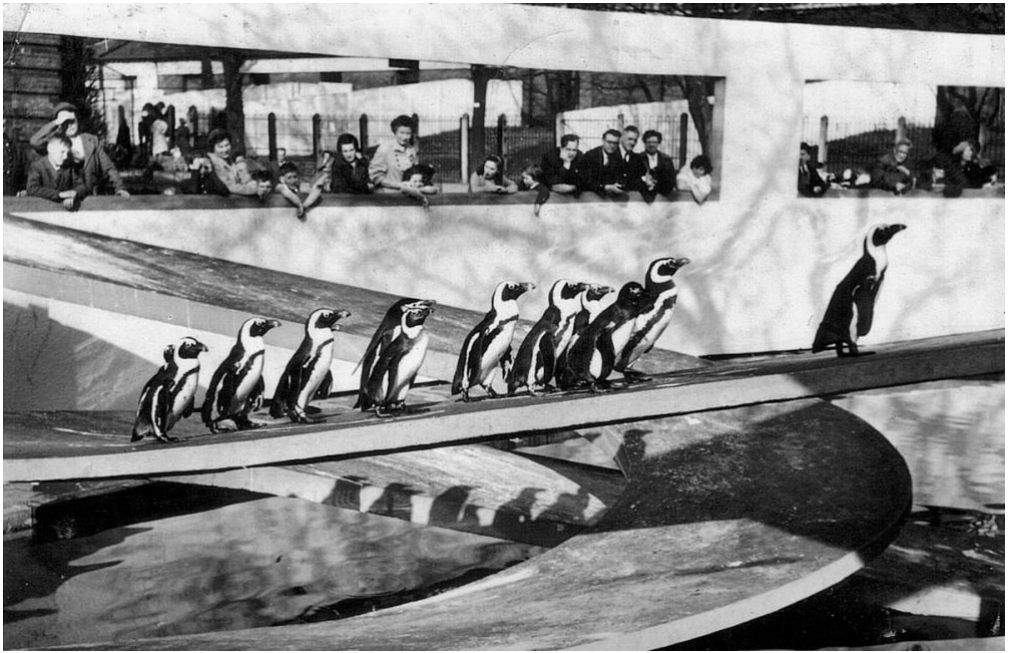
This paper explains why we need to go beyond

Evidential Pluralism accords with the weak thesis, not the strong thesis.



Causal inference requires a combination of statistical and mechanistic inference.

4 The replication crisis



When does a lack of replicability lead to a crisis?

That a study fails to replicate an observed association is a common occurrence.

This is normal science.

It is not of any concern in itself.

It is much more serious when established causal relationships are overturned.

An established proposition can be used as evidence for further claims.

A science makes progress to the extent that it establishes key propositions.

Causal propositions are key to most empirical sciences.

Regularly revoking established causal relationships leads to:

A lack of stability in the evidence base.

The need to re-evaluate claims made on the basis of revoked propositions.

Scepticism about the ability of a science to progress.

∴ A crisis arises when established causal relationships are often overturned.

IE When causal relationships are established on the basis of studies that fail to replicate.

This problem arises more regularly when:

1. We establish causation purely on the basis of association studies;
2. We appeal to few association studies;
3. The association studies are of lower quality;
4. Significance levels are too high;
5. There is publication bias and other kinds of bias;

   . . .

A lot of the action has centred around (4) and (5).

I would urge paying more attention to (1).

If Evidential Pluralism is correct:

Causal relationships are established by establishing mechanism as well as correlation.

We should scrutinise mechanistic studies alongside association studies.

The strong thesis is false.

A broader evidence base is more robust:

Biases are more easily eliminated.

It's less likely that an established claim hinges on a study that turns out not to be replicable.

A broad evidence base is arguably more important than successful replication.

Successful replication of a study may inspire false confidence in a causal claim.

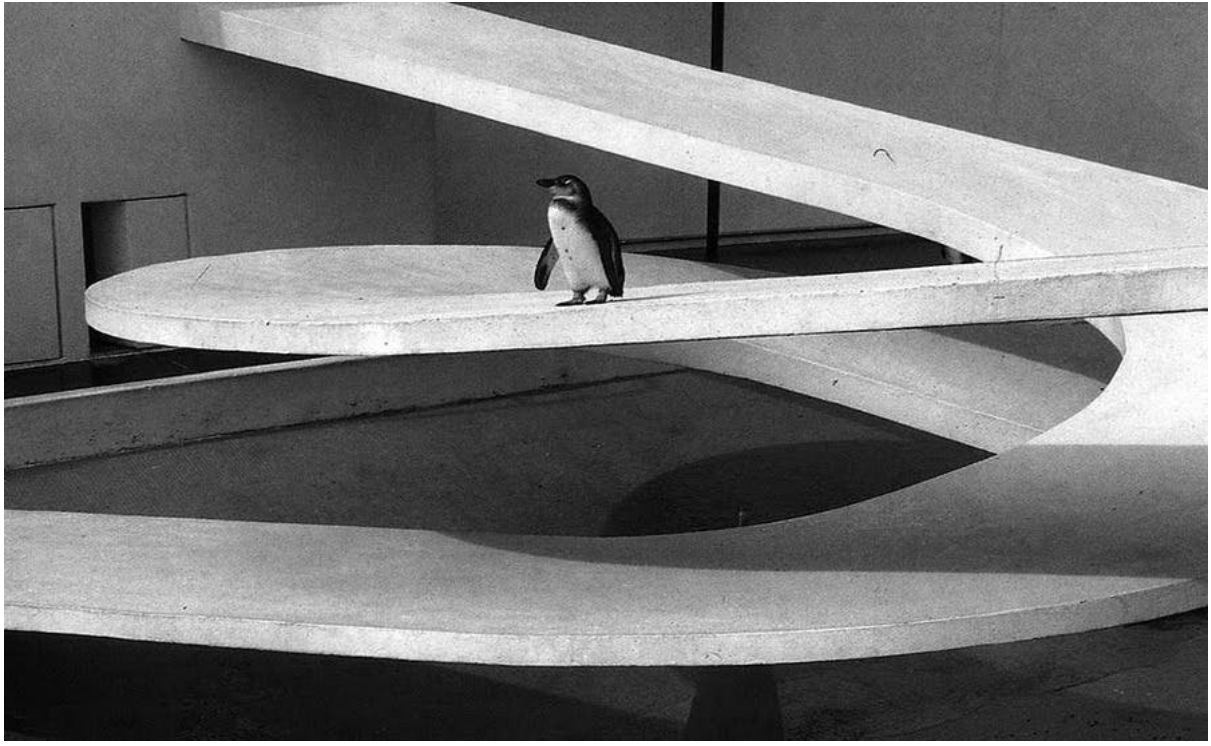
It may be that the biases of the study are replicated alongside the association.

Hence there is a need for triangulation (see, e.g., Munafò and Davey Smith, 2018).

Evidential Pluralism offers a well-motivated route to triangulation.

5 Examples

Four examples of Norton (2021, Chapter 3) show that Evidential Pluralism can lead to caution when establishing.



5.1 The discovery of Helicobacter pylori

In the early 1980s, Barry Marshall and Robin Warren found an association between the presence of Helicobacter pylori and gastritis and ulcers.

They submitted their paper to the Lancet in 1984.

> Reviewers didn't believe their findings.
>
> > There was evidence that bacteria could not survive in such an acidic environment.
> >
> > This negative mechanistic evidence undermined the association study.
>
> ∴ Their hypothesis was initially not considered established.
>
> So they contacted someone (Skirrow) who managed to replicate their findings.
>
> Only then did the Lancet publish.

By the late 1980s, there were:

> Many more association studies that replicated the findings,
>
> Mechanistic studies that found that *H*. pylori can survive in an acidic environment.

At this stage the hypothesis was generally regarded as established.

In the 1990s *H*. pylori was found to *require* an acidic environment (Clyne et al., 1995), and the mechanisms behind its survival are now better understood (Ansari and Yamaoka, 2017).

In 2005 Marshall and Warren won the Nobel prize for medicine for their work.

5.2 Cold fusion

In 1989 Martin Fleischmann and *B*. Stanley Pons announced that that had carried out fusion on a lab bench at ordinary temperatures.

> They electrolysed heavy water (deuterium oxide) with palladium electrodes.
>
> The cathode became saturated with deuterium.
>
> Large quantities of heat were produced.
>
> They argued that the deuterium atoms were brought close enough to ignite a nuclear fusion reaction.
>
> IE That fusion caused the large quantities of heat.

Steven Jones then announced that he had similar results.

However, the claim that these results were attributable to cold fusion was not established.

The US Energy Research Advisory Board argued in 1989 that:

Mechanistic evidence undermined a link between the heat and a nuclear process:

There were insufficient levels of fusion products (e.g., neutrons, tritium and isotopes of helium) for fusion to account for the heat.

Confirmed theory implied that electrostatic repulsion prevents free deuterium nuclei from approaching closer than about 0.1 nanometres, which is too far to initiate fusion.

The closest approach of deuterium atoms in palladium was found to be 0.17nm.

In contrast, in molecular deuterium, two deuterium atoms are 0.074nm apart.

Further attempts to replicate the experiment led to mixed results.

Replication attempts continued to lead to mixed results.

Similar effects have been observed many times, but the effects weren't routinely achieved.

A certain amount of luck is required to find these effects.

Edmund Sturms and others claimed that the successful replications demonstrate cold fusion.

The establishment consensus remains that the observed effects are not established to be attributable to cold fusion.

This is validated by Evidential Pluralism.

5.3 The Miller Experiment

The Michelson-Morley experiment of 1887 failed to detect an ether wind.

NB Ether is the supposed medium that carries light, electric fields and magnetic fields.

It should flow past the earth as it moves, if it exists. This is the ether wind.

This experiment confirms special relativity.

In 1925, Dayton C. Miller reported that an experiment found the ether wind.

This represented a failure of replication.

NB Miller was president of the American Physical Society.

His experimental apparatus was one of the most sensitive.

Einstein attributed Miller's result to experimental error related to temperature, because:

> Miller's assumption that the velocity of light is dependent on height above sea level is theoretically implausible.
>
> Another similar replication, the Trouton-Noble experiment, found no evidence of an ether wind.

In addition, Miller measured a flow of 10km/sec, but one would expect the motion of the earth would yield an ether wind flow of about 30km/sec.

Evidential Pluralism would condone scepticism here.

In 1955, Shankland et al. reanalysed Miller's results and found that they were associated with temperature variations in the apparatus.

> This analysis confirmed Einstein's doubts.

5.4 Intercessionary prayer

There have been several studies of the effectiveness of intercessionary prayer.

EG Leibovici (2001) conducted an RCT that found an association between remote, retroactive intercessionary prayer (in 2000) and length of stay of patients with blood infections in hospital (in 1990-6).

Byrd (1988) and Harris et al. (1999) also found positive associations between prayer and recovery of cardiac patients.

In general, however, association studies have yielded mixed results.

On the other hand, a systematic review of 10 studies found 'Specific complications (cardiac arrest, major surgery before discharge, need for a monitoring catheter in the heart) were significantly more likely to occur among those in the group not receiving prayer' (Roberts et al., 2009, p. 2).

Again, this is a case where well-confirmed scientific theory provides plenty of evidence against a mechanism.

Particularly in the case of retroactive prayer.

∴ Evidential Pluralism would not take the effectiveness of prayer to be established.

NB This accords well with Leibovici's own conclusions.

Summary

We used to agree that causation is not correlation.

- This apparently no longer the case, because:
  - The correlations and statistical techniques are getting more sophisticated.
  - There are more accounts that conflate causality with some kind of correlation.

This trend is pernicious.

- It leads to a simplistic account of causal enquiry.
- It is turning the normal phenomenon of replication failure into a crisis.

Evidential Pluralism provides a more comprehensive account of causal enquiry.

A broader evidence base is more robust in the face of potential replication failures.

Bibliography

Ansari, S. and Yamaoka, Y. (2017). Survival of Helicobacter pylori in gastric acidic territory. *Helicobacter*, 22(e12386).

Clyne, M., Labigne, A., and Drumm, B. (1995). Helicobacter pylori requires an acidic environment to survive in the presence of urea. *Infection and Immunity*, 63(5):1669–1673.

Fisher, R. (1935). *The design of experiments*. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1925). *Statistical methods for research workers*. Hafner, New York, thirteenth (1963) edition.

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

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

Hill, A. B. (1952). The clinical trial. *New England Journal of Medicine*, 247(4):113–119.

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of*

the Royal Society of Medicine, 58:295–300.
Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press, Oxford.
Leibovici, L. (2001). Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. *British Medical Journal*, 323:1450–1451.
Munafò, M. R. and Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature*, 553(7689):399–401.
Norton, J. D. (2021). *The Material Theory of Induction*. University of Calgary Press, Calgary.
Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley, Chichester.
Roberts, L., Ahmed, I., Hall, S., and Davison, A. (2009). Intercessory prayer for the alleviation of ill health. *The Cochrane database of systematic reviews*, 2009(2):CD000368.
Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.
Williamson, J. (2019). Evidential Proximity, Independence, and the evaluation of carcinogenicity. *Journal of Evaluation in Clinical Practice*, 25(6):955–961.