Hsiang-Ke Chao • Julian Reiss
Editors

Philosophy of Science in Practice

Nancy Cartwright and the Nature of Scientific Reasoning

Springer

Editors
Hsiang-Ke Chao
Department of Economics
National Tsing Hua University
Hsinchu, Taiwan

Julian Reiss
Department of Philosophy and Centre
for Humanities Engaging Science
and Society (CHESS)
Durham University
Durham, UK

Synthese Library
ISBN 978-3-319-45530-3 ISBN 978-3-319-45532-7 (eBook)
DOI 10.1007/978-3-319-45532-7

Library of Congress Control Number: 2016955047

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part III Models in Practice

On the Causal Wars

Julian Reiss

Very funny. Accused me of unscientific deduction.
'Whose expression is that?' I said to him.
'It's not deduction at all,' he says, 'it's induction.'
'My dear Percy, wherever have you been learning words like that:

you sound just like a beastly doctor or someone.' John Le Carré, Tinker Tailor Soldier Spy

1 Introduction

The 'causal wars' (Scriven 2008) are controversies currently plaguing a variety of fields of the applied social and life sciences including ageing studies, criminology, dentistry, development economics, education, medicine, psychology and social policy, among others. They are cold wars, but fought with intensity and some bitterness on both sides. Proponents of evidence-based approaches to scientific research, policy and practice are gaining ground at the minute, but no end is in sight.

The main bone of contention between the warring parties is the role randomised experiments should play in evaluations of claims about the efficacy of treatments and policies. Proponents on one side – sometimes referred to as 'randomistas' (e.g., Ravallion 2009) – maintain that randomised experiments are necessary for credible causal inference or at least play the role of a gold standard for judging the quality of evidence for efficacy claims. Their opponents – to whom we'll refer as 'inferentialists' for reasons that will become plain momentarily – not only deny randomised experimentation this special role (and *a fortiori* the idea that it should be necessary for credible inference) but draw attention to a variety of characteristics of the method that may make it inferior to some of its alternatives.

When one looks at the debate, one gets the impression that proponents of evidence-based approaches follow what one might call an 'epistemic paradigm'.

J. Reiss (✉)
Department of Philosophy and Centre for Humanities Engaging Science and Society (CHESS), Durham University, Durham, UK
e-mail: julian.reiss@durham.ac.uk

H.-K. Chao, J. Reiss (eds.), *Philosophy of Science in Practice*,
Synthese Library 379, DOI 10.1007/978-3-319-45532-7_4

Well-designed and executed randomised experiments are regarded as intrinsically reliable (or 'credible') and all other sources of evidence (such as observational studies or clinical judgement) have to live up to this standard. Subscribing to this paradigm has enormous rhetorical advantages. It is, first, extremely simple. One can explain what randomised experiments are to a novice in the course of just a few minutes and he or she will see that they're bound to work. Second, experimentation is widely understood to be the very essence of scientific method. It was when we stopped deriving scientific claims from first principles and unsystematic observation and started to experiment systematically when, or so the popular lore goes, the scientific revolution changed the way in which we investigate nature and society for ever. We would not want to lose the advantages experiments brought. Third, the 'randomistas' somehow managed to reserve the label 'evidence-based' for their approach/paradigm. This is unfortunate for anyone wishing to criticise them as, presumably, no-one would want to promote scientific research that refuses to be based on 'the best available evidence'.

Critics of randomised experiments do not have a unified paradigm. Some of the more technical work on randomisation is Bayesian in origin, but it is neither the case that all critics are Bayesian nor that the criticisms can only be rationalised from a Bayesian point of view. The arguments against randomised experimentation may well be convincing in their own right, but they are not motivated by a framework that presents a systematic alternative.

The aim of this paper is to sketch such a systematic alternative. I will start by briefly surveying the 'causal wars' in the three fields medicine, social science and economics. Then I'll introduce two alternative approaches to reasoning from evidence in the biomedical sciences and argue that one alternative has received little philosophical attention thus far. I will then proceed by articulating that alternative. This will form the core of the present paper. Finally I will come back to the causal wars, and show why randomised experiments appear attractive even from within the framework presented here but also show that the framework makes plain that they are neither necessary nor guaranteed to produce credible results.

2 The Causal Wars

For the benefit of readers who are not so familiar with the debates concerning evidence and causality currently going on in the philosophy of science and the affected sciences themselves, this section will briefly review some of the history of 'evidence-based' approaches in the social and biomedical sciences and some of the arguments that have been given on both sides. I will start by looking at medicine, where the evidence-based movement began, and then proceed to the social sciences and finally economics. Readers familiar with these debates should skip ahead.

2.1 Evidence-Based Medicine (EBM)

The 'evidence-based' label first appeared in clinical epidemiology and was used to describe a new approach to diagnostic, therapeutic and prognostic techniques in the management of patients developed at McMasters University in Hamilton, Ontario, by David Sackett, Gordon Guyatt, Brian Haynes and others. Sackett and his colleagues had observed that actual clinical practice does not always follow the recommendations of the latest clinical research but is instead, to an extent deemed inappropriate by the proponents of EBM, based on traditional wisdom, clinical experience and judgement. To the extent that the two conflict and that clinical research is more reliable, this can have deleterious consequences and indeed cost many lives and loss of well-being.

RCT evidence does sometimes conflict with evidence from other sources. John Worrall discusses three examples: grommets for glue ear, ventricular ectopic beats repressing substances such as encainide or flecainide for cardiac arrest, and routine foetal heart rate monitoring to prevent infant death (Worrall 2007, p. 985). In each case we have a procedure whose effectiveness is indicated by common sense and knowledge about the patho-physiological pathways – glue ear is a condition produced by a build-up of fluid in the middle ear that is unable to drain away because of pressure differentials, grommets act by letting air into the middle ear and thereby equalising pressure, for instance – but which, when tested by a randomised trial, turns out to be ineffective at best and positively harmful in the worst case. Cases such as these motivated Sackett et al. to call for a new approach to clinical medicine that 'de-emphasises intuition, unsystematic clinical expertise, and patho-physiologic rationale . . . [instead] and stresses the examination of evidence from clinical research' (Evidence-Based Medicine Working Group 1992, p. 2420).

So far, there is little with which one might possibly disagree. Things become more messy when we turn to EBM's understanding of 'evidence from clinical research'. When proponents of EBM speak of 'evidence' or 'systematic evidence' for the efficacy or effectiveness[1] of treatments, they refer to results produced by randomised experiments, or meta-analyses or systematic reviews of such results: 'It is when asking questions about therapy that we should try to avoid the non-experimental approaches, since these routinely lead to false positive conclusions about efficacy. Because the randomised trial, and especially the systematic review of several randomised trials, is so much more likely to inform us and so much less likely to mislead us, it has become the "gold standard" for judging whether a treatment does more good than harm' (Sackett et al. 1996).

Not everybody agrees that when results from RCTs and results from other sources of evidence diverge, the former should be preferred. A famous controversy in which the results from observational studies and those from RCTs conflicted was

[1] In medical research, 'efficacy' refers to the treatment's ability to affect the outcome in the desired direction in the trial population, 'effectiveness' to that ability in clinical practice (which means that it works also in different populations).

Level of evidence	Type of evidence
1++	High quality meta-analyses, systematic reviews of RCTs (including cluster RCTs), or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
1–*	Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias
2++	High quality systematic reviews of, or individual high quality non-randomised intervention studies (controlled non-randomised trial, controlled before-and-after, interrupted time series), comparative cohort and correlation studies with a very low risk of confounding, bias or chance
2+	Well conducted, non-randomised intervention studies (controlled non-randomised trial, controlled before-and-after, interrupted time series), comparative cohort and correlation studies with a low risk of confounding, bias or chance
2–*	Non-randomised intervention studies (controlled non-randomised trial, controlled before-and-after, interrupted time series), comparative cohort and correlation studies with a high risk of confounding, bias or chance
3	Non-analytical studies (eg case reports, case series)
4	Expert opinion, formal consensus

Fig. 1 A hierarchy of evidence (Source: Weightman et al. 2005)

that over the benefits and safety of hormone replacement therapy (HRT) in the early 2000s (Vandenbroucke 2009). HRT seemed protective for coronary heart disease in observational studies, whereas RCTs indicated an increase in the first years of use. For breast cancer, combined hormone preparations showed a smaller risk in an RCT than in observational studies.

In the end it turned out that the timescale of the effects was responsible, and that because of the way they are typically run, observational studies got some issues right and RCTs others (ibid.: 1234):

> The observational studies had picked up a true signal for the women closer to menopause. In the randomised trial, that signal was diluted because fewer women close to menopause were enrolled… The randomised trials had it right for coronary heart disease but failed to sufficiently focus on women close to menopause for breast cancer. The main reasons for the discrepancies were changes of the effects of HRT over different times…

Nevertheless, proponents of EBM insist on lexicographically ordered hierarchies of evidence that rank (systematic analyses of) RCTs ahead of (systematic analyses of) observational studies and all other sources of evidence. Figure 1 contains a typical example of such a hierarchy. What is striking is that it ranks even RCTs 'with a high risk of bias' ahead of 'high-quality non-randomised intervention studies'. Other hierarchies may differ in some of the details but they all agree on the relative importance of RCT evidence vs all other kinds of evidence.

Beginning as a small movement at McMaster University, Hamilton, Ontario, EBM is now a world-wide phenomenon and manifested in journals, editorial

practices and books, and institutionalised in international collaborations (such as the Cochrane Collaboration) and government agencies (such as NICE[2] or SIGN[3]). The latter means that EBM now has no small influence on how public health care decisions are made.

Critics argue that none of the considerations defenders of EBM present in favour of RCTs demonstrate that RCTs should play a special role in medical epistemology: it is not true that the logic of significance testing requires randomisation[4]; randomisation does not control for all confounders; it is at best one method to control for selection bias, among alternatives; we cannot without circularity assume that RCTs are always more reliable than observational studies; RCTs may 'clinch' their results, but only under a set of unrealistic assumptions and they are not the only method that operates in this way; in the best case, RCTs give evidence for efficacy but not for effectiveness – the quantity we're ultimately after – and the inference from former to latter is a hard one to make; they are narrow in their range of application.[5]

2.2 Social Science

Evidence-based approaches originated in biomedical research but we now find them all over the place in the social sciences. There are, for instance, evidence-based ageing and business studies, criminology, education research, information systems and library studies, management, public and social policy. Defenders of these approaches frequently appeal to medicine when they motivate and justify their views:

> Just as randomized trials revolutionized medicine in the twentieth century, they have the potential to revolutionize social policy during the twenty-first. (Duflo and Kremer 2008, p. 117)

> Randomised controlled trials (RCTs) are the best way of determining whether a policy is working. They have been used for over 60 years to compare the effectiveness of new medicines. […] This paper argues that we should and could use RCTs much more extensively in domestic public policy to test the effectiveness of new and existing interventions and variations thereof; to learn what is working and what is not; and to adapt our policies so that they steadily improve and evolve both in terms of quality and effectiveness. (Haynes et al. 2012, p. 6)

Now, there may be any number of things wrong with such an appeal to authority: perhaps medicine isn't a good model for social science because there are relevant

[2] National Institute for Health and Care Excellence, a non-departmental public body of the Department of Health in the United Kingdom.

[3] Scottish Intercollegiate Guidelines Network, a unit within Healthcare Improvement Scotland, itself a public body within the Scottish National Health System.

[4] Haavelmo (1944) can be read as an early statement of this point.

[5] The first four of these claims are made, for instance, by Worrall (2007), the next three, for instance, by Cartwright (2007).

differences[6]; perhaps there has never been a 'revolution' in medicine[7]; perhaps there has been, but only to the detriment of the quality of research results.[8]

Be that as it may, evidence-based social science is on the way to becoming as wide-spread and as influential as their counterpart in biomedical research. There are now six 'What Works Centres' in the UK whose goal is the systematic review of evidence for practitioners (i.e., policy makers) on the basis of the principles of evidence-based approaches. Similar institutions exist throughout the English-speaking world and elsewhere.

Unlike the controversy in medicine, which is, to my knowledge, fought mainly between defenders of evidence-based approaches from within the profession and critics who, by and large, are outsiders, in social science it is easy to find participants on both sides of the debates within each discipline. It is not surprising perhaps to find some critics to be associated with 'qualitative' approaches to social research. One argument one frequently hears from this camp is that RCTs draw attention away from trying to understand the complex causal structures that make efficacy claims true (an understanding which is said to be required to infer an effectiveness claim from an efficacy claim, for instance; see Byrne et al. 2009).

2.3 Economics

Economics gets a separate treatment here not because economics isn't a social science but because economics is characterised by methodologies that are quite distinct from those of other social sciences (or at least used to be quite distinct before other social sciences started to mimic economics more and more), because proponents of RCTs in economics do not call what they do 'evidence-based economics',[9] and, most importantly, because the approach is broader and therefore different than the 'RCT only' view we find in EBM and evidence-based social science.

---

[6] Here is one: in medicine, RCTs are one effective means to blind experimental subjects, thus helping to prevent various biases such as selection bias (Worrall 2002). Blinding is, however, seldom an option in the social sciences. It is possible to make placebos that look just like the new drug under test but it is not possible to hide from someone that they've been given a bed net for free, subsidised or at cost. So a major advantage of RCTs in medicine does not seem to apply in the social sciences (Scriven 2008).

[7] This is at least suggested by the following opener of Worrall's 2002 paper: 'The usual reaction from outside observers on being told that there is a (relatively) new movement called "Evidence-Based Medicine" is "What on earth was medicine based on before?" Telling clinicians that they ought to operate in accordance with EBM sounds about as controversial as telling people that they ought to operate in accordance with virtue.' (Worrall 2002, p. S316).

[8] There is evidence that the current state of biomedical research is not beyond reproach, to put it mildly (Als-Nielsen et al. 2003; Ioannidis 2005; Begley and Ellis 2012). As much of the criticised research was produced by RCTs, we have at least some reason to believe that the revolution (if it happened) wasn't necessarily for the better.

[9] This label has been used by Reiss (2004, 2008) as well as Joffe (2014) but refers to a completely different set of ideas.

Economists have realised that randomisation is in fact an instrumental variable in the sense econometricians use this term (Heckman 1996a); that is, a variable that causes the independent variable, causes the dependent variable, if at all, only through the dependent variable, and is not caused by the dependent variable or shares a common cause with it (Reiss 2005). Thus, instruments are the broader category, and randomisation is just one way to implement an instrument. Consequently, in economics we find the use of randomisation in field experiments in development economics (Duflo et al. 2004) but also the use of government elections as an instrument to estimate the effect of hiring policemen on crime (Levitt 1997), the use the Fed's error in forecasting of the Treasury account balance to measure the liquidity effect (Hamilton 1997) and a two-state least squares analysis (which is one way to compute an instrumental-variable estimate) of a natural experiment concerning minimum wages (Card and Krueger 1994). Randomised field experiments, instrumental-variable studies, two-stage least squares and similar techniques are sometimes jointly referred to as 'design-based econometrics' within empirical economics (Angrist and Pischke 2010).

In econometrics, the critics of 'design-based' approaches come, once more not too surprisingly, from defenders of structural econometrics in the tradition of the Cowles Commission. One frequently made argument is that an instrument is very likely to be invalid unless backed up by a theoretical model. Heckman (1996b), for instance, criticises Angrist's (1990) use of the random sequence number RSN (which had been used to determine draft status for cohorts of young men in the Vietnam War era) to estimate the effect of serving in the war on civil earnings later in life. Heckman argues that in a behavioural model employers would observe that number, use it to infer the probability of an employee to be drafted and consequently make investment in training decisions dependent on it (which would invalidate the RSN as an instrument). Another argument that is to my knowledge original to economics is that risk averse subjects who are aware that they are part of a randomised experiments and therefore made to play a lottery may leave the experiment, making the experimental population unrepresentative of the more general population researchers would like to study (Heckman 1992).

3 *Two* Approaches to Reasoning from Evidence

Despite the success of evidence-based medicine, it would be a mistake to assume that biomedical research is characterised by a single approach to reasoning from evidence. Instead, there are two alternative and competing approaches. There is, on the one hand, the experimentalist approach, according to which randomised experiments are the gold standard of evidence for causal claims in biomedical research. It did not, however, originate with the evidence-based medicine movement in the 1980s. It is much older, going back to at least Ronald Fisher's work in biostatistics of the 1920s and 1930s. On the other hand, there is the inferentialist approach, which maintains that causal claims are established by judgements based on diverse

bodies of evidence using pragmatic criteria. This approach played an important role, for instance, in the smoking/lung cancer controversy of the 1950s (Parascandola 2004).

The experimentalist approach has received a great deal of philosophical attention and defence since the times of Francis Bacon. Indeed, Mill's method of difference can be understood as a formalisation of the confirmation of a causal claim by a controlled experiment. RCTs are nothing but a probabilistic version of this. Inferentialism has received much less attention. In particular, it plays no role in the recent debate concerning evidence-based medicine in philosophy of science. And yet, countless widely accepted causal claims have never been subjected to an experiment. Smoking and lung cancer is one case in point. John Worrall mentions penicillin in the treatment of pneumonia, aspirin for mild headache, diuretics for heart failure, appendectomy for acute appendicitis and cholecystectomy for gallstone disease (Worrall 2007, p. 986); Jeremy Howick adds automatic external defibrillation to start a stopped heart, tracheostomy to open a blocked air passage, the Heimlich manoeuvre to dislodge an obstruction in the breathing passages, rabies vaccines and epinephrine in the treatment of anaphylactic shock (Howick 2011, p. 40). We probably know more causal claims that haven't been tested experimentally than we know claims that have.

One reason for the lack of interest in the inferentialist approach may be that it is a lot harder to articulate. It raises a number of difficult issues:

- What is a 'body of evidence'?
- Just how 'diverse' does it have to be?
- What are the 'pragmatic criteria' according to which we're supposed to come to a judgement about the hypothesis?
- How precisely are judgements made?

In what follows I will present my own proposal for addressing these issues.

4 The Framework

The framework I am developing for the articulation of the inferentialist approach to reasoning from evidence is called *eliminativist hypothetico-contextualism* (EHC). It has, thus, three dimensions that I will introduce in order of familiarity, starting with the most familiar one.

4.1 Hypothetico-Deductivism and Its Discontents

The hypothetico-deductive theory of evidence, according to which an observation statement e is evidence for a hypothesis h if and only if e can be deduced from h in conjunction with background knowledge b (so that $h\&b\vdash e$),[10] has long since fallen out of favour among philosophers of science. As early as 1980, Clark Glymour called the theory 'hopeless' (Glymour 1980). This stands in stark contrast to many scientists across the biomedical and social sciences who seem to regard hypothetico-deductivism as correctly describing evidential reasoning in their discipline (e.g., Deaton 2010; Gurven 2006; Hopayian 2004).[11]

Philosophers have given up on the theory because it suffers from countless riddles and counterexamples they believe damage it beyond repair (e.g., Hempel 1945). Let me focus on only two issues here.[12] First, typical scientific hypotheses entail very little about possible observations, even in conjunction with background knowledge. Second, the fact that h entails e is entirely consistent with there being alternative hypotheses h', h'' etc. which also entail e. Now, as long as we're only interested in evidence in some absolute sense, i.e., in asking whether the observation is consistent with or symptomatic of the truth of our hypothesis, this is fine. But evidence also plays the role of confirming hypotheses, and confirmation is always relative to alternatives (often simply thought of as the negation of the hypothesis).[13] The ability of h to entail e does not guarantee that e is also evidence in this relative, discriminatory sense. In the remainder of this section I will address the first problem, reserving the next section for the second one.

Hypotheses entail very little about possible observations we might make, no matter what and how much background knowledge we add. To begin with a mundane example, consider a murder case. There is no doubt that a person's, say, Smith's, fingerprints (or other traces) on the murder weapon are evidence for the hypothesis 'Smith murdered Jones'. But if Smith did murder Jones, there's no guarantee that he left finger prints. He may have worn gloves; he may have wiped off the murder weapon; he may have thrown it in a river; it may have rained; a cat may have licked them off; he may have a genetic defect preventing the formation of friction ridges and so on.

---

[10] As Pierre Duhem taught us over a century ago, scientific hypotheses are never tested in isolation. He maintained that for instance theoretical hypotheses are tested in 'bundles' that also include auxiliary hypotheses about the experimental apparatus and its proper functioning, background theories etc. (Duhem 1991 [1914]).

[11] For some evidence concerning the hypothesis driven nature of research in biomechanical and biomedical research, see Rowbottom and Alexander (2012) and Haufe (2013).

[12] I will ignore more technical problems such as the ravens paradox and tacking-on paradoxes here. I believe, however, that they do not arise in the contextualist version of the theory I propose. The demonstration will have to wait for another paper.

[13] For those who believe that hypotheses and evidential statements come with associated probabilities, this is easy to see: $P(e)$ decomposes into $P(h)P(e, h) + P(\neg h)P(e, \neg h)$, and $P(e)$ plays a role in all the standard confirmation/corroboration functions discussed in the literature.

Now, it is important to note that adding background knowledge is not in any way helpful here. First, even if we knew that Smith did not wear gloves, that he had nothing to wipe off the weapon, that there wasn't a river nearby, that it didn't rain and so on, there's still no guarantee that there will be fingerprints because there are indefinitely many ways in which fingerprints can be prevented from being left and remaining on a murder weapon until the forensics come in.

Second, we obviously never have this kind of knowledge. We don't know whether or not Smith wore gloves or whether a cat came by to lick off his fingerprints. Even in the best case we can only remove some possibilities. There are other possibilities we might consider but are not in the position to verify and yet other possibilities we would never think of (Rumsfeld's infamous 'unknown unknowns'). None of this stops fingerprints from being evidence for claims about a suspect's guilt.

The scientific case is analogous. Correlations (whether recorded in an observational or experimental study) are evidence for causal hypotheses. There is no guarantee, however, that if an independent variable I causes a dependent variable D, that I and D will be correlated in the data sets that have been recorded. At the causal level, it may be the case that I causes D through two mechanisms, one of which influences D positively, one negatively, so that the overall effect cancels out to zero.[14] Or perhaps cancellation isn't the problem but mismeasurement is. Or inadvertent conditioning on a mediating variable through data selection. And so on.[15]

[14] Even if we accept arguments to the effect that such exact cancellations have Lebesgue measure zero, the problem remains to distinguish between small correlations that are due to chance (and do not have a causal explanation) and those that are due to cancellation (and do have a causal explanation). From the point of view of empirical research, the argument about the near impossibility of exact cancellations cuts no ice.

[15] Darrell Rowbottom (personal communication) has pointed out that whether or not adding background propositions solves the problem depends on what exactly the background propositions are – 'knowledge' (the term I have been using), 'information' (a term he seems to favour, see Rowbottom 2014) or 'assumptions' (e.g., Williamson 2015). First, let me point out that the notion of 'background knowledge' I have in mind is not of the 'justified true belief' sort traditional epistemologists favour. Chances are that in many cases a scientist will rely on justified true beliefs when forming expectations, but counterexamples can readily be constructed. A beginning PhD student may employ regression analysis because she was instructed to do so by her advisor without actually *believing* that correlations are evidence for causal relations. Whether or not background propositions need to be justified depends on the model of justification one favours. On my own contextualist view of justification (Reiss 2014, 2015), a researcher will be justified in taking certain propositions as given, but not under standard externalist models (because the results of an unreliable method can nevertheless be used to ground expectations in a given case) or standard internalist models (because the researcher may not have access to the justification). In many cases, background propositions may later turn out to be false. I would instead cash out the notion as one of 'material background commitments' a researcher is contextually justified in maintaining. I'd be happy to refer to these as 'assumptions' except that researchers are contextually justified in making some assumptions but not others. Importantly, researchers will seldom if ever be contextually justified in making all those background assumptions needed to derive an evidential statement from the hypothesis and the assumptions. 'Perpetrator killed Victim' might entail 'Perpetrator's fingerprints are on the weapon' together with 'Perpetrator used weapon to kill Victim', 'Perpetrator did not wear gloves', 'Perpetrator has ridges on fingers', 'Perpetrator thoroughly touched weapon with her naked hand to kill Victim' and 'Nothing between the killing and the analysis of the weapon hap-

The (admittedly fairly obvious) key to a solution to this problem is to think of the relation between *h* and *e* not in deductive but in inductive terms. *h* does not entail *e* but rather entitles us to expect *e* to be the case.[16] Given what we know about how the world works, we know that murderers sometimes leave fingerprints on murder weapons – and murder weapons leave traces on murderers. So we're entitled to expect to find fingerprints under the supposition of the hypothesis. Importantly, not to find them means as good as nothing on its own. Traces can disappear for any number of reasons. Similarly, causal relations tend to produce certain patterns in the data – correlations, regularities, invariant relationships, traces of processes and so on. Thus, under the supposition of the hypothesis, we expect to find these patterns in the data (without drawing important conclusions when we don't find them).

This inductive relation between hypothesis and evidence is similar to the probabilistic notion of partial entailment. *h* partially entails *e* if and only if $P(h \mid e) > P(h)$. The difference is that I do not presuppose here that there are such probabilities or that they represent anything meaningful beyond arbitrary degrees of belief. There are certainly no *physical* probabilities of murderers leaving fingerprints (not to mention probabilities that given fingerprints were left, they remained intact until the forensics investigated the murder weapon).[17] When dactyloscopy was first developed for the identification of individuals in the late eighteenth century, we would expect murders to have been a lot less careful than they might be today. Today whether or not fingerprints are likely to be left has to do with the degree to which the act was planned, and it would be preposterous to assume that there are stable probability measures over these plans. Nor are there measures over all the different ways in which fingerprints, once left, may disappear.

We can of course always come up with some number to describe subjective degrees of belief. However, unless they are based in the physical facts (which they aren't), they must be regarded as arbitrary and pretending a degree of precision that does not exist. The right logic for evidential reasoning is a logic of possibility, not of probability. If Smith did murder Jones, then he *may* have left fingerprints on the weapon, just as we *may* find gunshot residues on him. But the probability P(there are fingerprints on the weapon|Smith murdered Jones) is undefined.

---

pened that removed Perpetrator's fingerprints', but I cannot imagine a scenario in which these assumptions can plausibly be made.

[16] Note that my account implies is that the *h-e* relation is an *a posteriori* one, as Achinstein (1995) demands it should be.

[17] I believe this to be true independently of whether we understand physical probabilities along propensity or relative frequency lines. The relative frequency interpretation would be somewhat more plausible in the present context, but I maintain with Hacking (1965) and others that these probabilities arise relative to 'chance set-ups', i.e., relatively stable special kinds of arrangements of features that are triggered or 'run' repeatedly so that meaningful relative frequencies can be computed. The set-up within which killers leave fingerprints on their weapons isn't of this kind.

4.2 Eliminativism

The second problem of the standard hypothetico-deductive theory was that a piece of evidence *e*, qua following from *h*, does not speak in favour of *h* over and above alternatives to *h*, from which *e* follows equally. This problem can be solved by demanding that evidence of two kinds be collected: direct evidence, which pertains to the hypothesis of interest *h*; and indirect evidence, which serves to rule out alternatives. We can define:

e is *direct evidence* for a hypothesis *h* if and only if *e* is a pattern in the data we are entitled to expect under the supposition of *h*.
e is *indirect evidence* for a hypothesis *h* if and only if *e* is a pattern in the data that is incompatible with what we are entitled to expect under the supposition of any of h's alternatives h', h'' etc.

Direct evidence speaks in favour of the hypothesis by showing that what we'd expect to be the case were the hypothesis true is actually the case. Indirect evidence speaks in favour of the hypothesis by showing that alternatives cannot account for all known facts. Suppose Jones was stabbed with a 10 in. bread knife. Smith's fingerprints on it belong to the direct evidence that Smith killed Jones. An alternative account for the presence of Smith's fingerprints would be, for instance, the hypothesis that the knife is Smith's, he regularly uses it to slice bread, and so we are entitled to expect his fingerprints on it even if the original hypothesis were false. However, if we now learn that Smith has a cleaner who comes every Tuesday and puts all knives in the dish washer (which, we suppose, removes all fingerprints), and the killing happened just after she left Smith's house, this new fact is incompatible with the alternative 'Smith left these fingerprints when slicing bread some time in the past'.

Something analogous happens when causal hypotheses are established in the sciences. If *I* is hypothesised to cause *D*, we're entitled to expect to find the two variables correlated. But a host of alternative hypotheses can equally account for a correlation between *I* and *D*: '*D* causes *I*', 'A common factor *Z* causes both *I* and *D*', '*I* or *D* or both have been mismeasured', '*I* and *D* jointly cause common effect *E*, and the population from which the current data were drawn is conditioned on *E*' and so on. Facts about the temporal order of the variables can, for instance, rule out reverse causation. A new study that conditions on all potential common causes and finds that the correlation persists rules out the common-cause hypothesis. And so on.

However, these two modifications – to regard the relationship between hypothesis and evidence as inductive rather than deductive and to require indirect evidence which rules out alternative hypotheses – seem to create more problems than they solve:

- How do we determine what we are entitled to expect?
- How do we determine whether given alternatives (of which there may be indefinitely many) are *relevant* and stand in need of consideration/elimination?

- How much evidence is needed in order to eliminate an alternative?
- How does one end the regress of having to show that alternative accounts of the evidence at hand are false?

It is at this point that context plays important roles. These will be considered in the next subsection.

4.3 Contextualism

'Context' is in this paper understood as having three dimensions:

- Background knowledge
- The nature and purpose of the inquiry
- Normative commitments.

Let us consider these in turn.

Background Knowledge Background knowledge from hundreds of years of history of science tells us how causal factors operate – at a highly general level across all domains of inquiry, at the level of a scientific discipline such as medicine and at a more specific field within a discipline such as cancer research. Until well into the twentieth century 'causality' and 'determinism' were conceptually inseparable. Causes were thought to necessitate their effects (or, to use a more Humean language, to be invariably followed by their effects). Even if, after Mill's work on causation in the nineteenth century, it was understood that causes are often complex sets of conditions and we might not know all the members of the set that are needed for the effect to invariably follow, for every effect it remained supposed that there is such a set of conditions. This changed only with the enormous empirical success of quantum mechanics but remained controversial for a long time after that. Today most scientists and philosophers of science believe some version of the probabilistic theory of causation according to which even the full set of causal conditions may only raise the probability of the effect without making it inevitable to be more adequate. This is fully consistent with the existence of deterministic relations, but these relations are a special case of the more general category 'causal relations' which also include indeterministic relations.

In the nineteenth century, a view about medical causation according to which disease causes are (among other things) necessary conditions for the onset of a disease was extraordinarily successful and led to the discovery of countless micro organisms such as viruses, bacteria, protists, fungi and prions that are responsible for diseases. It was soon discovered, however, that many diseases do not have a 'characteristic' germ that is found in all cases of the disease. Most if not all cancers

can be brought about by a variety of different factors, and so the search for necessary conditions appears to be in vain.[18]

Background knowledge of this kind helps to determine which patterns in the data to expect and which not to expect. Under a deterministic conception of disease causation, we can expect the disease to develop when the micro organism is introduced into a healthy organism. This is in fact the third postulate of the so-called 'Koch's Postulates'. Koch himself found later that there were asymptomatic cholera patients (Koch 1893) and therefore abandoned this criterion. Here, then, we have a *bona fide* cause of a disease, the bacterium *Vibrio cholerae* which acts in ways that are inconsistent with what we would expect of causes. We therefore learn that disease agents, at least as far as infectious diseases are concerned, are not always sufficient for their effects. And this in turn tells us that a potential disease agent that is associated with but not always followed a disease should not be eliminated from the 'list of suspects'.

Background knowledge also tells us about alternatives. It's a simple mathematical fact that two independent (or uncorrelated) variables, conditioned on a joint effect, can be correlated. That retrospective studies in epidemiological research often suffer from this problem is a fact about how the data are drawn. When they are drawn exclusively from hospitalised patients, a correlation between two variables I and D can be accounted for by selection bias, which is due to the fact that both I and D cause patients to be hospitalised. We therefore know that when data are drawn from hospitalised patients, selection bias is a relevant alternative that must be eliminated.

Nature and Purpose of the Study It is a methodological injunction not to consider certain outré alternatives in scientific inquiries. Of course it is conceivable that an evil demon is responsible for a correlation between two variables and all 'scientific' alternatives are false. However, if someone were to invoke evil-demon hypotheses he would change the nature of the inquiry (cf. Williams 2001). A scientific inquiry would be replaced by a philosophical one. Or, to put it in a different way, in a philosophy class room, especially when fundamental principles of epistemology are being discussed, evil-demon hypotheses are relevant. They are never relevant in a scientific inquiry.

The purpose of the inquiry helps to make a causal hypothesis more precise and thereby provides information about what direct evidence to collect and which alternatives are relevant. To find out whether administering I or or bringing I about through policy is an effective strategy to change D does not necessarily require knowledge of the mechanism through which I causes D. That lithium is effective in the treatment of mania has been known since the 1940s but the exact mechanism is still not understood. Often researchers have also explanatory interests, however, and

[18] Cervical cancer may constitute a counterexample as it is caused in well over 90 % by a human papilloma virus (HPV) (Walboomers et al. 1999). But even in this case the situation is rather complicated as there are many different types of HPV, and cancer risk differs dramatically among the different types.

the 'black-box' causal claim that administering lithium helps to improve mania symptoms does not go very far in satisfying these interests.

The reverse can also be the case. If '*I* causes *D*' has been established by means of finding a causal chain '*I* causes M_1 causes M_2 causes… causes M_n causes *D*' there is no guarantee that bringing about *I* will lead to a change in *D*. On the one hand, it may be the case that all possible interventions on *I* have a preventer of at least one M_i as a side effect. On the other hand, the causal chain might be extremely fragile so that any changes in the causal background (including changes that have *I* as an effect) disrupt the chain so that upon intervention, *I* no longer causes *D*. Knowing this causal hypothesis may well be useful for explaining present and past outcomes but not for building effective strategies.

The purposes of an inquiry will also be informative of the risks involved in regarding a hypothesis as established or refraining from doing so. The potential costs of accepting 'Smoking causes lung cancer' when it is in fact false or failing to accept it when it is true are vastly higher than those of, say, accepting or rejecting a hypothesis about the causes of Queen Elizabeth's use of 'etc.' in stating her full title (this is Ernest Nagel's example, see Nagel 1979, p. 552ff.). This then leads us to the norms characterising an inquiry.

Normative Commitments There are two senses in which the EHC framework as presented here may lead to a never ending inquiry. There is, first, the consideration of alternatives, of which there may be infinitely many. When our primary evidence is a retrospective study, it will almost always be good idea to rule out selection bias. But alternatives may become increasingly outré. By this I mean not philosopher's thought experiments but genuine empirical possibilities that nevertheless seem far fetched.

To give a fictional example, in the TV sitcom *Frasier*, the main character Dr Frasier Crane, a radio psychiatrist, tries to help his father Martin, a retired police detective, solve an old murder case. Martin mulls over the case in a number of episodes. Frasier finally takes pity and a good look at the case, and feels confident that his degrees from Harvard and Oxford will compensate for his lack of experience in the field (especially his deficiency relative to the father). He then comes up with the hypothesis that Brisbane, one of the suspects of the case and a man who could teach his monkey to impersonate George Washington, trained the monkey to cock a revolver, sneak up the fire escape, lie await for Helen, the victim, pump her full of lead and then make his getaway perhaps even still wearing the revolutionary war regalia in order to confuse any chance witnesses. Is this possibility too outlandish to be considered?

In the sitcom, the hypothesis was invoked to poke fun at Frasier's hubris. In another fictional example, an initially no less implausible hypothesis was introduced to show the reverse: that even highly unlikely alternatives can turn out to be true. The example is from Arthur Conan Doyle's *Silver Blaze* (Doyle 1894) where it eventually turned out that a horse was Straker's 'murderer'. It is hypotheses such as this to which Doyle referred when he let his Sherlock Holmes say, 'when you have eliminated the impossible, whatever remains, however improbable, must be the

truth' (Doyle 1890).[19] One should add that 'impossible and 'improbable' are notions whose conditions of application vary with context. If the last remaining alternative is a hypothesis describing aliens descending from a UFO, abducting the victim, inflicting the fatal wound and returning him to Earth, then we probably had better reconsider the 'impossible' scenarios – those that have already been eliminated.

The second way in which an inquiry can potentially continue for ever is because it involves an infinite regress. Any indirect evidence that helps to rule out an alternative can itself be accounted for by alternative hypotheses. A police detective, quite obviously, never observes whether a suspect's alibi is true. All he has is evidence – the suspect says she was with her girlfriend at the time of the crime, the girlfriend confirms this, they produce a credit card bill for a meal for two, paid for with the suspect's card. But of course, the suspect may have lied in order to protect herself, the girlfriend covers her and went to the restaurant with someone else using the suspect's credit card. So we can go on and ask the waiter if he remembers the suspect and girlfriend having dinner; but he too can lie, misremember the girlfriend's dinner companion, bear a grudge against the police for reasons unrelated to the case and make something up just to annoy the detectives, or, most realistically, simply not care one way or another. And so on, ad infinitum.

Scientific inquiry does not come for free. There are direct, opportunity and ethical costs. These costs have to be traded off against the benefits of reducing uncertainty. The benefits of reducing uncertainty consist in the reduced chance of accepting a false or rejecting a true hypothesis. In many Western countries, it is a principle of criminal law that a suspect is found guilty only if one is pretty sure that she committed the crime. This means that also relatively bizarre alternative hypotheses should be considered and eliminated. Similarly, indirect evidence that serves to rule out alternatives should be probed relatively deeply.

In science, things tend not to be so clear-cut. If 'smoking causes lung cancer' is accepted though it is in fact false, many people will be worried needlessly, stop smoking and forgo enjoyment, might be stigmatised for nothing and so on. When it is rejected although true, there will be considerable health costs.[20] To determine what is more important a value judgement is necessary.

Value judgements such as this will then help to eliminate alternatives. An important alternative in the smoking/lung cancer controversy of the 1950s was diagnostic error: the observed correlation between smoking and lung cancer incidence was supposed not to have been genuine but rather an artefact of mistaking deaths due to other conditions (for instance, tuberculosis) for deaths from lung cancer. This alternative was eventually ruled out by a study that looked at age and gender patterns in mortality data (Gilliam 1955). For instance, assuming that lung cancer prevalence was stable over time would mean diagnostic error of only 3 % among those 35–44 years of age but 59 % among those 75 years or older as well as different rates

[19] I'm not sure to what extent it is relevant that Doyle was trained and worked as a physician. He wrote his dissertation on syphilitic myelopathy. Doyle, apparently, modelled Sherlock Holmes on his university teacher Joseph Bell (Chalmers 2006).

[20] It has been argued that there are also benefits because people who die of lung cancer die a relatively cheap death and are likely to die a much more costly death, say of Alzheimer's, later (Heath 2001).

for men and women. Now this is of course possible and could be explained by alternative hypotheses. Perhaps coroners are ageists and look a lot more carefully in determining the cause of death when the body is that of a young person than when it is that of an older person. But if, and this is what has to be judged, the costs of maintaining this very implausible alternative even though it's false are much higher than eliminating it even though it's true, which is at least plausible in the smoking/lung cancer case, then we have a good reason to consider it eliminated.

5 Inferential Judgement

How, then, should we answer the questions concerning the inferentialist approach that were raised above in Sect. 2? Let us go through them in turn.

'Body of Evidence' The body of evidence is simply given by the direct evidence plus the indirect evidence that pertains to all relevant alternatives.

'Diverse' The diversity of the body of evidence is related to the fact that the alternative hypotheses are usually a very diverse bunch. It requires quite different patterns in the data to rule out, say, that a correlation is in fact due to a common cause, than to rule out that it was produced by a programming error or outright fraud. Conditioning on common causes (if common causes are known and measurable) helps in the former case, replication in the latter. Alternatively, a study that shows that peer review is very effective in detecting programming errors and fraud in the field at hand would rule out the latter but not be relevant to the former.

'Pragmatic Criteria' Numerous pragmatic criteria have been discussed in the previous section. Let me add a few here. Scientific research has a 'default and challenge' structure (Brandom 1994; Williams 2001). What I mean by that is that researchers are entitled to other researchers' results unless there a specific reason not to trust them. If I know that of 53 published 'landmark studies' in preclinical cancer research, the results of only 6 (11 %) could be replicated (Begley and Ellis 2012), I might look more closely before building my own study on the results of others (in that area).

Other pragmatic criteria include paying attention to effect size and study characteristics. If effects are large, and the contributions of other factors known at least approximately, it may be possible to rule out that the other factors account for the entire effect (essentially by applying Mill's method of residues). This is also something that happened in the smoking/lung cancer controversy. Smokers, especially strong smokers, have a dramatically increased risk of contracting lung cancer. There may be an unmeasured common cause that is responsible for smoking as well as lung cancer, but known genetic factors were able to account for risks of much smaller sizes. Finally, the study I mentioned above that stratified mortality data with respect to age and gender used an extremely simple study design. A study like this is a lot less likely to be error prone than studies using complex statistical methods or experimental designs or any other forms of high inferential leverage.

'Judgements' Judgements are, essentially, mappings from the evidence in a context to a proposition of the form 'There is weak/moderate/strong evidence/proof that *h*'. What is important to note is that there are no universal rules. There are very often very good reasons to judge one way rather than another, but reasons don't determine outcomes.

To make this plain, consider three issues from the smoking/lung cancer controversy. First, while it is clear that some alternatives such as section bias or diagnostic error are always relevant alternatives in this type of research, and evil-demon hypotheses never, the relevance of some alternatives is contentious. Consider the common-cause hypothesis. Of course it is possible that there is a genetic factor that both makes people want to smoke and causes cancer. Fisher even had some evidence that monozygotic twins are more concordant with respect to smoking behaviour than dizygotic twins, which speaks in favour of smoking having a genetic basis. However, given that lung cancer rates very suspiciously shot up about a generation or so after people exchanged smoking pipes for smoking cigarettes, the genetic hypothesis or, as it was known then, the 'constitutional hypothesis', would require a very sudden and very large scale mutation in a large part of the population. Moreover, as women started to smoke cigarettes in large numbers some years after men did, and, lo and behold, lung cancer rates shot up again about a generation later, the mutation must have acted in highly unconventional ways. Now this is certainly implausible. But at what 'degree of implausibility' should one consider a hypothesis irrelevant (or eliminated by the facts)?

A similar continuum exists when we look at how effect size helps to eliminate alternatives. As is well known, there is a large dose-response curve in the smoking-lung cancer risk. Moderate smokers have an about ninefold increased risk, strong smokers about 60-fold. At the time of the controversy it was known that blood type, which itself was known to be genetically based, was implicated in cancer susceptibility. Blood type could, however, account for only 20 % of the risk. So if smoking increased lung cancer risk only 20 % and not 800–5,900 %, the genetic hypothesis could clearly not be ruled out on the basis of this consideration. Perhaps it is as clear that with the observed effect size the genetic hypothesis could incontrovertibly be ruled out.[21] But where to draw the line?

Lastly, let me mention a conceptual issue. Above in Sect. 3.3. I argued that background knowledge for instance about how causes operate in certain domains helps to determine what kinds of facts to collect as evidence for hypotheses. But to call what we have learned about causal relations 'knowledge' may be a bit of an exaggeration. Jacob Henle, a nineteenth century German physician, made the following remarks about causes in medicine (Henle 1844, p. 25; quoted from Carter 2003, p. 24):

> Only in medicine are there causes that have hundreds of consequences or that can, on arbitrary occasions, remain entirely without effect. Only in medicine can the same effect flow from the most varied possible sources. […] This is just as scientific as if a physicist were to teach that bodies fall because boards or beams are removed, because ropes or cables break, or because of openings, and so forth.

[21] Some notable participants in the historical debate disagreed!

The kind of thinking about disease causation Henle describes here has led to the adoption of the germ theory of disease that was extraordinarily successful in the nineteenth century, as was noted above. But today it is regarded as mistaken. Cancers usually have multiple causes, and each of these causes usually has multiple effects. Lung cancer is caused by smoking, but also by inhaling asbestos, radon gas, other pollutants, and genetically. Smoking, in turn, causes lung cancer, but also coronary heart disease, stroke, all sorts of other cancers, low birth weight and many more deficiencies. We may describe the current state of the art in cancer research using Henle's exact same words.

Trying to find causes that were the necessary universal conditions for a disease was successful in the nineteenth century because it led to the discovery of a host of micro organisms which allowed hypothesising about transmission channels, which in turn allowed the development of effective strategies to prevent transmission. Now, it may well be that there are necessary conditions in carcinogenesis, perhaps certain kinds of mutations that are characteristic of certain kinds of cancer. However, given the current state of knowledge and technological abilities, this kind of knowledge is not very helpful for devising therapies or prevention strategies. It's a lot easier to stop smoking than to stop a certain mutation from happening in a body's cell. It is therefore that today environmental and behavioural factors are regarded as causes, with the consequence that effects have multiple causes and causes have multiple effects.

However, it is at least conceivable that 200 years or so down the line our knowledge and technical abilities to have increased enough so that therapies can target mutations directly. If so, we may well return to an age of necessary universal causes (or something similar).

Now, the point is that there is no objective fact about what should be regarded as a cause and what not. To be a cause means to be useful in explanation, prediction, control. But what is useful in explanation, prediction, control depends on our explanatory interests, practical goals and technical and financial abilities. Thus, if someone rules out some hypothesis about a cause of cancer on the basis that the factor at stake is associated also with a large number of other diseases and conditions (as Joseph Berkson did), he may well reason anachronistically, but I would hesitate in calling him 'objectively wrong'.

Thus, drawing a conclusion about a hypothesis of interest requires facts, normative considerations, but also a number of judgements that could easily go either way. Therefore I say that judgements about hypotheses are based on reasons, but reasons don't determine these judgements.

6 Back to the Causal Wars

Finally, let us apply some of these considerations to the causal wars. On the one hand, the inferentialist approach, as articulated here, explains what proponents of RCTs find so attractive, perhaps even compelling, about well-designed and

executed RCTs: they rule out a number of salient alternative hypotheses in one fell swoop. If treatment and control group are balanced with respect to all other causes of the outcome and the treatment is correlated with the outcome, then common-cause and reverse causation hypotheses can indeed be ruled out. But the inferentialist approach also makes clear that randomised experimentation is in no way necessary to achieve this result. Any experimental design, for instance, allows to rule out reverse causation.[22] How Fisher's constitutional hypothesis was ruled out by purely observational means has been discussed above.

More importantly, any experimental result can be accounted for by hypotheses that have nothing to do with the experimental design as such. Variables must be measured properly; correlations computed correctly; results reported errorlessly; data pooled adequately and so on. Alternative hypotheses describing these kinds of errors (that, for instance, a correlation is due to diagnostic error rather than a causal relation) have to be ruled out by non-experimental means anyway. An experimental design, therefore, at best allows the ruling out of *some* salient alternatives, the alternative causal hypotheses. Other alternatives have to be eliminated by other means, and it should therefore be a small step to allow these other means to eliminate the salient causal alternatives as well.

7 Conclusions

My goal here was to sketch an alternative to the currently dominant experimentalist paradigm in evidence-based medicine, policy and practice. The alternative is a development of the hypothetico-deductive theory of evidence, suitably modified to avoid the major drawbacks of the standard version of this theory.

Within the inferentialist approach controlled experiments come out as *one way* to eliminate salient alternatives (but not more), and it's possible to have highly credible evidence from observational data. The big issue that remains is of course whether these considerations will sound convincing to experimentalists who, it seems to me, have a fundamentally different, foundationalist, epistemology. To defend contextualism against foundationalism is, alas, a topic for a different paper.

References

Achinstein, P. (1995). Are empirical evidence claims a priori? *British Journal for Philosophy of Science, 46*(4), 447–473.

Als-Nielsen, B., Chen, W., Gluud, C., & Kjaergard, L. (2003). Association of funding and conclusion in randomized drug trials: A reflection of treatment effect of adverse events? *Journal of the American Medical Association, 290*(7), 921–928.

[22] If the treatment or policy is administered by the experimenter, then it should not be correlated with the outcome unless the treatment or policy causes the outcome.

Angrist, J. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security adminstrative records. *American Economic Review, 80*(3), 313–336.
Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives, 24*(2), 3–30.
Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531–533.
Brandom, R. (1994). *Making it explicit: Reasoning, representing & discursive commitment*. Cambridge, MA: Harvard University Press.
Byrne, D., Olsen, W., & Duggan, S. (2009). Causality and interpretation in qualitative policy related research. In D. Byrne & C. Ragin (Eds.), *Handbook of case-centred research* (pp. 511–521). London: Sage.
Card, D., & Krueger, A. (1994). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review, 84*(4), 772–793.
Carter, K. C. (2003). *The rise of causal concepts of disease: Case histories*. Aldershot: Ashgate.
Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*(2), 11–20.
Chalmers, J. (2006, August 6). Conan Doyle and Joseph Bell: The real Sherlock Holmes. In *The independent* (online edition). Edinburgh: Surgeons' Hall Museums.
Deaton, A. (2010). Understanding the mechanisms of economic development. *Journal of Economic Perspectives, 24*(3), 3–16.
Doyle, A. C. (1890). *The sign of four*. London: Spencer Blackett.
Doyle, A. C. (1894). Silver blaze. In: *The memoirs of Sherlock Holmes*. London: Newnes.
Duflo, E., & Kremer, M. (2008). Use of randomization in the evaluation of development. In W. Easterly (Ed.), *Reinventing foreign aid* (pp. 93–120). Cambridge, MA: The MIT Press.
Duflo, E., Glennerster, R., & Kremer, M. (2004). Randomized evaluations of interventions in social service delivery. *Development Outreach, 6*(1), 26–29.
Duhem, P. (1991 [1914]). *The aim and structure of physical theory*. Princeton: Princeton University Press.
Evidence-Based Medicine Working Group. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268*(17), 2420–2425.
Gilliam, A. (1955). Trends of mortality attributed to carcinoma of the lung: Possible effects of faulty certification of deaths due to other respiratory diseases. *Cancer, 8*, 1130–1136.
Glymour, C. (1980). Discussion: Hypothetico-deductivism is hopeless. *Philosophy of Science, 47*, 322–325.
Gurven, M. (2006). Human behavioral ecology. In H. Birx (Ed.), *Encyclopedia of anthropology*. Thousand Oaks: SAGE.
Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, *12*(Supplement), iii–vi and 1–115.
Hacking, I. (1965). *The logic of statistical inference*. Cambridge: Cambridge University Press.
Hamilton, J. (1997). Measuring the liquidity effect. *American Economic Review, 87*(1), 80–97.
Haufe, C. (2013). Why do funding agencies favor hypothesis testing? *Studies in History and Philosophy of Science Part A, 44*(3), 363–374.
Haynes, L., Service, O., Goldacre, B., & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. London: Cabinet Office- Behavioural Insights Team.
Heath, J. (2001). *The efficient society*. London: Penguin.
Heckman, J. (1992). Randomization and social policy evaluation. In C. F. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs* (pp. 201–230). Cambridge, MA: Harvard University Press.
Heckman, J. (1996a). Randomization as an instrumental variable. *The Review of Economics and Statistics, 78*(2), 336–341.
Heckman, J. (1996b). Comment. *Journal of the American Statistical Association, 91*(434), 459–462.
Hempel, C. (1945). Studies in the logic of confirmation (I.). *Mind, 54*(213), 1–26.

Henle, J. (1844). Medicinische Wissenschaft und Empirie. *Zeitschrift für rationelle Medizin, 1*, 1–35.
Hopayian, K. (2004). Why medicine still needs a scientific foundation: Restating the hypotheticodeductive model – Part two. *British Journal of General Practice, 54*(502), 402–403.
Howick, J. (2011). *The philosophy of evidence-based medicine*. Oxford: Wiley-Blackwell.
Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.
Joffe, M. (2014, April). Can economics be evidence-based? *Royal Economic Society Newsletter*, *2014*(165).
Koch, R. (1893). Über den augenblicklichen Stand der bakteriologischen Choleradiagnose. *Zeitschrift für Hygiene und Infectionskrankheiten, 14*, 319–333.
Levitt, S. (1997). Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review, 87*(3), 270–290.
Nagel, E. (1979). *The structure of science*. Indianapolis: Hackett.
Parascandola, M. (2004). Two approaches to etiology: The debate over smoking and lung cancer in the 1950s. *Endeavour, 28*(2), 81–86.
Ravallion, M. (2009). Should the randomistas rule? *The Economists' Voice, 6*(2), 1–5.
Reiss, J. (2004). Evidence-based economics: Some questions and preliminary answers. *Analyse and Kritik, 26*, 346–363.
Reiss, J. (2005). Causal instrumental variables and interventions. *Philosophy of Science, 72*(PSA 2004), 964.
Reiss, J. (2008). *Error in economics: Towards a more evidence-based methodology*. London: Routledge.
Reiss, J. (2014). What's wrong with our theories of evidence? *Theoria, 29*(2), 283–306.
Reiss, J. (2015). A pragmatist theory of evidence. *Philosophy of Science, 82*(3), 341–362.
Rowbottom, D. (2014). Information versus knowledge in confirmation theory. *Logique et Analyse, 57*(226), 137–149.
Rowbottom, D., & Alexander, N. (2012). The role of hypotheses in biomechanical research. *Science in Context, 25*(2), 247–262.
Sackett, D., Rosenberg, W., Gray, M., Haynes, B., & Richardson, S. (1996). Evidence-based medicine: What it is and what it isn't. *British Medical Journal, 312*, 71–72.
Scriven, M. (2008). A summative evaluation of RCT methodology & an alternative approach to causal research. *Journal of MultiDisciplinary Evaluation, 5*(9), 11–24.
Vandenbroucke, J. P. (2009). The HRT controversy: Observational studies and RCTs fall in line. *The Lancet, 373*, 1233–1235.
Walboomers, J., Jacobs, M., Manos, M., Bosch, X., Kummer, A., Shah, K., Snijders, P., Peto, J., Meijer, C., & Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *Journal of Pathology, 189*(1), 12–19.
Weightman, A., Ellis, S., Cullum, A., Sander, L., & Turley, R. (2005). *Grading evidence and recommendations for public health interventions: Developing and piloting a framework*. London: Health Development Agency.
Williams, M. (2001). *Problems of knowledge*. Oxford: Oxford University Press.
Williamson, J. (2015). Deliberation, judgement and the nature of evidence. *Economics and Philosophy, 31*(1), 27–65.
Worrall, J. (2002). What evidence in evidence-based medicine. *Philosophy of Science, 69*, S316–S330.
Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass, 2*, 981–1022.

Julian Reiss is Professor of Philosophy at Durham University and Co-Director of the Centre for Humanities Engaging Science and Society (CHESS). He is the author of *Error in Economics* (2008), *Philosophy of Economics* (2013), *Causation, Evidence, and Inference* (2015), and over 50 journal articles and book chapters on topics in the philosophy of the biomedical and social sciences.