⋆

Mechanisms and the Evidence Hierarchy

⋆

Brendan Clarke[a], Donald Gillies[a], Phyllis Illari[a],
Federica Russo[b] and Jon Williamson[c]

[a] Science and Technology Studies, University College London, UK
[a] Dipartimento di Studi Umanistici, Università degli Studi di Ferrara, Italy
[c] Philosophy, SECL, University of Kent, UK

Draft of October 22, 2013
To appear in *Topoi*, special issue on
'Evidence and Causality in the Sciences'

Abstract

Evidence-based medicine (EBM) makes use of explicit procedures for grading evidence for causal claims. Normally, these procedures categorise evidence of correlation produced by statistical trials as better evidence for a causal claim than evidence of mechanisms produced by other methods. We argue, in contrast, that evidence of mechanisms needs to be viewed as complementary to, rather than inferior to, evidence of correlation. In this paper we first set out the case for treating evidence of mechanisms alongside evidence of correlation in explicit protocols for evaluating evidence. Next we provide case studies which exemplify the ways in which evidence of mechanisms complements evidence of correlation in practice. Finally, we put forward some general considerations as to how the two sorts of evidence can be more closely integrated by EBM.

§1 Introduction

Sackett et al. (1996) characterise evidence-based medicine (EBM) as follows:

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients.

In order to make decisions about patient care, one typically needs to diagnose—to determine the most probable cause of the patient's symptoms—and treat—to determine which treatment intervention is most likely to alleviate the diagnosed causes. Thus one needs to establish what causes what and one needs to apply this causal knowledge to new patients. This paper is concerned with methods for establishing and using causal claims, particularly in evidence-based medicine.

The EBM movement has transformed the way in which evidence is gathered and evaluated in medicine. Medical researchers and those charged with making

Level of evidence	Type of evidence
1^{++}	High-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1^{+}	Well-conducted meta-analyses, systematic reviews 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 case–control or cohort studies High-quality case–control or cohort studies with a very low risk of confounding, bias or chance and a high probability that the relationship is causal
2^{+}	Well-conducted case–control or cohort studies with a low risk of confounding, bias or chance and a moderate probability that the relationship is causal
2^{-}	Case–control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is not causal [a]
3	Non-analytic studies (for example, case reports, case series)
4	Expert opinion, formal consensus
[a] Studies with a level of evidence '–' should not be used as a basis for making a recommendation (see section 7.4)	

Figure 1: Hierarchy of evidence for intervention studies from NICE (2006, p.47).

Levels of evidence	Type of evidence
Ia	Systematic review (with homogeneity)[a] of level-1 studies[b]
Ib	Level-1 studies[b]
II	Level-2 studies[c] Systematic reviews of level-2 studies
III	Level-3 studies[d] Systematic reviews of level-3 studies
IV	Consensus, expert committee reports or opinions and/or clinical experience without explicit critical appraisal; or based on physiology, bench research or 'first principles'

[a] Homogeneity means there are no or minor variations in the directions and degrees of results between individual studies that are included in the systematic review.

[b] Level-1 studies are studies:
- that use a blind comparison of the test with a validated reference standard (gold standard)
- in a sample of patients that reflects the population to whom the test would apply.

[c] Level-2 studies are studies that have **only one** of the following:
- narrow population (the sample does not reflect the population to whom the test would apply)
- use a poor reference standard (defined as that where the 'test' is included in the 'reference', or where the 'testing' affects the 'reference')
- the comparison between the test and reference standard is not blind
- case–control studies.

[d] Level-3 studies are studies that have **at least two or three** of the features listed for level-2 studies.

Figure 2: Hierarchy of evidence for diagnostic studies from NICE (2006, p.48).

Criteria for assigning grade of evidence	
Type of evidence	Randomized trial = high Observational study = low Any other evidence = very low
Decrease* grade if	• Serious or very serious limitation to study quality • Important inconsistency • Some or major uncertainty about directness • Imprecise or sparse data • High probability of reporting bias
Increase grade if	• Strong evidence of association—significant relative risk of > 2 (< 0.5) based on consistent evidence from two or more observational studies, with no plausible confounders (+1) • Very strong evidence of association—significant relative risk of > 5 (< 0.2) based on direct evidence with no major threats to validity (+2) • Evidence of a dose response gradient (+1) • All plausible confounders would have reduced the effect (+1)
Range	High quality evidence Moderate quality evidence Low quality evidence Very low quality evidence

Figure 3: The GRADE system advocated by NICE (2009). Source: http://www.gradeworkinggroup.org/FAQ/evidence_qual.htm.

treatment and public health decisions now tend to be guided by explicit *evidence hierarchies*. An evidence hierarchy ranks evidence for a causal claim. Fig. 1, for example, depicts an evidence hierarchy advocated by the UK National Institute for Health and Care Excellence for evaluating treatment effectiveness, while Fig. 2 is a corresponding hierarchy for evaluating diagnostic claims (NICE, 2006). More recently, NICE advocated the GRADE system depicted in Fig. 3, which highlights the main point of commonality between the plethora of evidence hierarchies that abound in the literature: randomised trials (RCTs) are ranked more highly than observational studies, which in turn are ranked more highly than any other kind of evidence, other things being equal. Evidence hierarchies have become entrenched in medicine and are now spreading to other areas, particularly to the social sciences and to public policy.

Evidence hierarchies have met with some controversy in the philosophical literature. The merits of randomised trials, and the question of whether they should trump other sorts of trial, have been thoroughly debated by Papineau (1994); La Caze (2008, 2009); La Caze et al. (2012); Cartwright (2010); Cartwright and Munro (2010); Northcott (2012); Worrall (2002, 2007, 2010), for example. Moreover, most hierarchies rank meta-analyses and systematic reviews more highly even than randomised trials, a move that has been criticised by Stegenga (2011). Thus the philosophical literature has been mostly concerned with the top end of the evidence hierarchies—i.e., with the way in which randomised trials and meta-analyses are exalted by these hierarchies.[1]

In contrast, this paper focuses on the bottom end of the evidence hierarchies. In this paper we raise the concern that evidence of mechanisms, which is normally

[1]We recognise that hierarchies are capable of playing other roles in facilitating medical decision making too, such as providing safeguards against litigation, or to simplify prescription practices (Timmermans and Berg, 2003). However, our focus here is firmly on their epistemological role.

Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
How common is the problem?	Local and current random sample surveys (or censuses)	Systematic review of surveys that allow matching to local circumstances**	Local non-random sample**	Case-series**	n/a
Is this diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards**	Case-control studies, or "poor or non-independent reference standard**	Mechanism-based reasoning
What will happen if we do not add a therapy? (Prognosis)	Systematic review of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial*	Case-series or case-control studies, or poor quality prognostic cohort study**	n/a
Does this intervention help? (Treatment Benefits)	Systematic review of randomized trials or *n*-of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study**	Case-series, case-control studies, or historically controlled studies**	Mechanism-based reasoning
What are the COMMON harms? (Treatment Harms)	Systematic review of randomized trials, systematic review of nested case-control studies, *n*-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning
What are the RARE harms? (Treatment Harms)	Systematic review of randomized trials or *n*-of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect			
Is this (early detection) test worthwhile? (Screening)	Systematic review of randomized trials	Randomized trial	Non -randomized controlled cohort/follow-up study**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning

Figure 4: Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence (OCEBM Levels of Evidence Working Group, 2011).

relegated to the bottom of the evidence hierarchies, should be treated alongside, rather than as inferior to, the evidence provided by statistical trials. (Statistical trials—such as randomised trials, cohort studies, case-control studies, case series and n of 1 trials—primarily test whether the putative cause is correlated with the putative effect, and, if so, how strong a correlation there is. In this paper, we use 'correlation' in the broad sense of probabilistic dependence between arbitrary variables, as opposed to, e.g., the narrow sense of a linear correlation coefficient of two continuous variables.) In particular, the paper has two aims. The first is to build on Russo and Williamson (2011a); Illari (2011); La Caze (2011) and Clarke et al. (2013) in setting out the case for treating evidence of mechanisms alongside evidence of correlation in medicine (§2). The second is to give some indication as to how this can be achieved (§3, §4).

Here we take a broad view of mechanisms: 'a mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon' (Illari and Williamson, 2012, p. 120). Evidence of mechanisms can take a wide variety of forms, ranging from laboratory experiments to literature reviews of basic science to individual patient case studies to textbook consensus to expert testimony. Clearly, evidence of mechanisms can vary in quality just as can evidence of correlation.

That evidence hierarchies take a dim view of evidence of mechanisms is witnessed by the fact that such evidence is normally restricted to the lowest level of a hierarchy, the other levels being reserved for evidence obtained from various sorts of statistical trial. This is perhaps clearest in the latest hierarchy produced by the Oxford Centre for Evidence-Based Medicine, where the lowest level, level 5, is reserved for 'mechanism-based reasoning' (Fig. 4).[2] Similarly, Fig. 2 places evidence based on physiology and bench research at its lowest level, while the GRADE system (Fig. 3)

[2] 'Mechanism-based reasoning' seems to refer roughly to what we call evidence of mechanisms. Howick writes 'Mechanistic reasoning is an inferential chain (or web) linking the intervention (such as HRT) with a patient-relevant outcome, via relevant mechanisms.' (Howick, 2011, p.929). We will address this point in more detail later.

grades evidence other than that obtained from a randomised trial or observational (i.e., non-interventional) study as of the lowest quality.

The plan of the paper is as follows. In §2 we present the case for considering evidence of mechanisms alongside evidence of correlation: evidence gleaned from randomised trials is fallible and there is room to consider other sorts of evidence alongside such evidence (§2.1); recent work in the philosophy of causality and the history of medicine suggests that in order to establish a causal claim one normally needs to establish both that the putative cause is correlated with the putative effect and that there exists some underlying mechanism that can account for this correlation (§2.2); evidence of mechanisms is required in order to adequately design and interpret randomised trials (§2.3); evidence of mechanisms is required in order to mitigate the problem of external validity (§2.4); evidence of mechanisms is required in order to apply a general causal claim to a specific individual (§2.5); but evidence of mechanisms has its own limitations, in particular a complexity problem and a masking problem, and should be used in conjunction with evidence of correlation (§2.6). In §3 we present a tuberculosis case study that illustrates the ways in which evidence of mechanisms can be used in conjunction with evidence of correlation in practice. In §4 we put forward some general guidelines for integrating the two kinds of evidence.

§2
Why integrate evidence of mechanisms and evidence of correlation?

In this section we argue that evidence-based medicine should integrate evidence of mechanisms with evidence of correlation because taken on their own each sort of evidence has significant limitations (§2.1, §2.6), and because they need to be taken together in order to establish causal claims (§2.2, §2.3), to transport causal claims to new populations (§2.4) and to apply causal claims to individual patients (§2.5).

§2.1. The fallibility of statistical trials

When evaluating the claim that variable A causes variable B, statistical trials may be performed to determine whether A and B are suitably *correlated*. Such trials consider individuals with differing values of A and determine the extent to which the value of B covaries with that of A. (It is not essential that several individuals are involved: an *n of 1* trial considers changes in A and B over time in a single individual. What is important in a statistical trial is that there is a large number of instantiations of A.) Typically, a statistical trial will be used to estimate the probability distribution $P(B|AC)$, where C is some variable capturing background factors—normally, known causes of B—which are controlled for or held fixed. If A and B are found to be probabilistically dependent conditional on C, then A and B are deemed suitably correlated in context C, and the causal claim is confirmed.

As with any statistical test, there are a number of ways in which the trial can mislead. For example, small sample sizes or sample bias can lead to sample correlations when there is no dependence between A and B in the population as a whole, or no sample correlation when there is a dependence, or can yield a poor estimate of the extent to which A and B are dependent in the population. But one problem is peculiar to causal inference, namely the problem of confounding: even if it is true that A and B are dependent conditional on known causes C, they may not be dependent conditional on *all* of B's causes—i.e., it may not be A that is making a difference to

B in the trial; the change in B may be being made by unknown 'confounding' causes C' of B which happen to covary with A.

Randomised trials are used to alleviate the problem of confounding. By randomly allocating the values of A in the trial, it is hoped that any dependence between A and unknown causes C' of B will be broken, so that, as the trial size increases, any observed correlation between A and B is less and less likely to be attributable to unknown causes. However, the attempted randomisation may fail to break the link between A and C'. There may remain a systematic connection between A and C', as is the case, for example, when patients can tell which treatment they are receiving, thereby inducing a differential placebo effect. Or there may be a coincidental correlation between A and C': a randomised trial is only likely to fully eradicate a correlation between A and C' in the asymptotic limit, as the number of allocations of A tends to infinity. In practice, a randomised trial can only involve relatively few allocations, leading to a realistic chance of a coincidental correlation between A and confounding causes (see, e.g., Thompson, 2011, §2.2). This realistic chance of stumbling across coincidental correlations contributes to an apparently paradoxical difference when reasoning about RCTs in theoretical and practical contexts: while an ideal RCT can theoretically provide warrant for a causal claim on its own (Cartwright, 2007, 63), most 'real' RCTs—even well-designed and perfectly executed ones—do not provide any such unequivocal warrant. Trials that do give strong support to a causal claim do so in virtue of specific expert judgement 'baked in' to their design, rather than as a consequence of their logical structure. As Cartwright puts it, "RCTs need a number of demanding assumptions beyond valid reasoning." (Cartwright, 2007, 68). If further evidence was needed of the exceptional nature of unequivocal RCTs, it can be found in the practice of terminating prematurely RCTs that show such results (Bassler et al, 2010).

The upshot is this. While evidence hierarchies may correctly identify the relative merits of different sorts of statistical trial, any such trial is very fallible. Hence it is by no means the case that, when evidence is available from one or more trials high up the hierarchy, one should ignore evidence from trials lower down the hierarchy or indeed non-statistical evidence of underlying mechanisms. This point is now fairly well recognised insofar as it applies to statistical trials. While some decision makers retain the view that trials higher up the hierarchy *trump* those lower down, and when available one should consider the former while ignoring the latter, most systematic reviews aggregate evidence from all statistical trials—those higher up the hierarchy may be given greater weight, but those lower down are usually not entirely ignored. The point is much less well recognised as it pertains to evidence of mechanisms. Normally, when there is evidence available from statistical trials, non-statistical evidence of mechanisms is simply ignored. This is for two main reasons. First, as we saw in §1, non-statistical evidence of mechanisms is often placed at the lowest level of the hierarchy of evidence, and the lowest level tends to be considered so poor-quality as to be trumped by evidence at higher levels. Second, it is hard to see how to systematically consider qualitative, non-statistical evidence of mechanisms alongside quantitative, statistical evidence of correlation when conducting systematic reviews and meta-analyses.[3]

[3]The International Agency for Research on Cancer (IARC) is one of the few agencies that now tries to systematically consider evidence of mechanisms (IARC, 2006, §B.4). One way it does this is by formulating a two dimensional hierarchy that considers evidence obtained in experimental studies on animals along one dimension and evidence obtained on humans along the other. However, on each dimension the emphasis is still on evidence obtained from statistical trials.

We shall argue next that one ought to integrate evidence of mechanisms with evidence of correlation when trying to establish a causal claim. In §4 we shall discuss systematic ways in which this can be achieved.

§2.2. The epistemology of causality

In this section we shall describe a recent line of work in the philosophy of causality that concerns the question of which evidence is needed to establish a causal claim.

Russo and Williamson (2007) argued in favour of the following epistemological thesis:

> *RWT.* In order to establish that A is a cause of B in medicine one normally needs to establish two things. First, that A and B are suitably correlated—typically, that A and B are probabilistically dependent, conditional on B's other known causes. Second, that there is some underlying mechanism linking A and B that can account for the difference that A makes to B.

Note that, according to this epistemological thesis, what is required is evidence of two different sorts of things—correlation and mechanisms—not two different *kinds* of evidence (Illari, 2011). Indeed, a single item of evidence can be evidence of both correlation and mechanisms. For instance, in principle a well devised and well conducted randomised trial can on its own provide evidence for a causal claim, since it can provide evidence of correlation, and, if in the circumstances other explanations of this correlation are sufficiently implausible, it can also provide evidence that there is some underlying mechanism linking the putative cause and the putative effect that can account for the correlation. (In practice, however, it is rare that a randomised trial is large enough and of high enough quality to *establish* a causal claim on its own.)

A variety of considerations support RWT, as we shall now see.

¶ *Non-causal correlations.* Evidence of an appropriate sort of correlation between A and B cannot be enough to establish a causal connection between A and B, because correlations can arise in a great variety of ways, only one of which is causal connection between A and B.

The problem of confounding is one illustration of this fact: the problem is that the correlation between A and B may be attributable to some other cause of B, rather than to A causing B. In such situations there can be evidence that is good enough to establish the appropriate sort of correlation, yet the hypothesis that this correlation is due to confounding may be more plausible than the hypothesis that it is due to some underlying mechanism. In such cases we would of course be reluctant to regard the causal claim as established. To give an extreme example, Leibovici (2001) provides good evidence from a randomised trial in favour of there being a correlation between remote, retroactive intercessionary prayer and length of stay of certain patients in hospital, although, the authors acknowledge, 'no mechanism known today can account for the effects of remote, retroactive intercessory prayer said for a group of patients with a bloodstream infection' (Leibovici, 2001, p. 1451). Other examples of good evidence of correlation in the absence of good evidence of mechanisms include studies in precognition (Bem, 2011) and homeopathy (Cucherat et al., 2000). Examples like these show that strong evidence of correlation is not sufficient to establish a causal claim.

But correlations can also arise in other ways. High bread prices in Britain are correlated with high sea levels in Venice, not because one causes the other or because of a common cause of each, but because bread prices in Britain and sea levels in Venice are both increasing due to largely independent mechanisms (Sober, 1988). Similarly, the prevalence of celiac disease is correlated with the global spread of HIV, simply because both are increasing for independent reasons. It is our mechanistic evidence—our evidence that there is no mechanistic connection between the variables in question that can account for the observed correlation—that prevents us from attributing these correlations to causal connections. (This is not of course to say that we should uncritically take evidence of such mechanisms at face value—see §4. The point is rather that evidence of mechanisms can be substantial enough and of high enough quality so as to override evidence of correlation.)

Correlations also arise when variables are semantically connected. The government might increase taxation of the unmarried in order to appropriate some of the disposable income of bachelors. This is an effective strategy not because taxation of the unmarried is a cause of taxation of bachelors. The two are not distinct and so do not stand in the cause-effect relation: taxation of bachelors *is* taxation of the unmarried. The correlation between taxation of the unmarried and taxation of bachelors is attributable to the semantic relation between 'bachelor' and 'unmarried', not to a causal connection. Such semantic relations are rife in medicine. For example, with the benefit of hindsight we now know that the correlation between cases of phthiasis, consumption and scrofula is attributable to a semantic connection, rather than a causal connection: all three terms refer to tuberculosis. It is advances in our knowledge of the mechanisms of disease that allows us to attribute such correlations correctly.

Correlations can also be attributable to logical connections (particularly between logically complex variables), physical connections (e.g., the law of conservation of total momentum) and mathematical connections (e.g., mean and variance variables are dependent in virtue of being defined relative to the same distribution)—see, e.g., Williamson (2005, §4.2) for discussion of such cases. Evidence of mechanisms can often help us distinguish these sorts of correlations from causal correlations.

Even in cases where a correlation between A and B is indeed attributable to a causal connection between the two variables, the correlation may be insufficient to establish a causal relation because it is not clear which variable is the cause and which is the effect. It is often evidence of the underlying mechanisms that allows us to differentiate between the two alternative causal claims and establish one of them. It is evidence of mechanisms, not evidence of correlation, that stops us from deeming the thermometer reading to be a cause of the temperature, or the presence of mud to be a cause of rain.

¶ *Medical Methodology.* The epistemological thesis RWT receives some support from writings on medical methodology. Before the advent of evidence hierarchies, the Bradford Hill criteria constituted the predominant guidelines for discovering causal relationships in medicine. Bradford Hill (1965) argued in favour of the following list of indicators of causality: (1) strength of association; (2) consistency of the observed association; (3) specificity of the association; (4) temporality (the cause occurs before the effect); (5) biological gradient (the dose-response curve); (6) biological plausibility; (7) coherence 'with the generally known facts of the natural history and biology of the disease'; (8) experimental evidence; (9) analogy ('with the effects of thalidomide and

rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy').

Broadly speaking, items 1,2,3,5,8 are strong indicators of correlation, while items 4,5,6,7,8,9 are strong indicators of an underlying mechanism. While Bradford Hill argues that none of these indicators is *necessary* for establishing causality, the fact that they provide a balance between evidence of correlation and evidence of mechanism does accord with RWT. (This balance dates back at least to Bernard (1856), who advocated a mixed methodology of statistical studies and physiological experimentation in medicine.) It is this balance, we would argue, that has been lost in present-day evidence hierarchies.

¶ *Instances of causal discovery.* The general epistemological thesis RWT also receives some support from past attempts to establish causal claims in medicine.

Classic examples of causal discovery seem to support the thesis (Russo and Williamson, 2007). Let us first consider the example of Koch's efforts to prevent cholera (Brock, 1988, 229-232). These were stimulated by a serious outbreak of cholera in Hamburg in 1892. Hamburg has a neighbouring city Altona further down the Elbe, but curiously Altona was nearly free of cholera. What made this more surprising was that Hamburg's sewage was carried down the Elbe to Altona. Just for this reason, however, Altona filtered its water supply using slow sand filters. Hamburg, however, did not filter its water. This evidence of correlation strongly suggested that slow sand filtration prevented cholera. However, this conclusion was not generally accepted and was, in particular, rejected by Koch's opponent Pettenkofer.

Koch had isolated the cholera vibrio in 1884, and suggested that it was the cause of cholera. Using this hypothesis, he now proposed a mechanism, namely that slow sand filtration removed the cholera vibrio. This mechanism could be tested out by bacterial counts before and after slow sand filtration. The results strongly confirmed the correctness of Koch's mechanism. When this evidence of mechanism was added to the earlier evidence of correlation, Koch's view became generally accepted, and was adopted by the German government in its efforts to prevent further cholera outbreaks.

In this cholera example, the evidence of correlation occurred first; the causal claim was only later clinched by evidence of a mechanism. However, the opposite order is to be found in another classic case—the discovery of the cause of anthrax (Debré, 1994, 294-318, 378-413). In the first three quarters of the 19th century, anthrax was a very serious disease of cattle and sheep, which was sometimes contracted by humans as well. Davaine, a French scientist, suggested in 1863 that the disease was caused by a micro-organism which he called 'bacteridium', and which is now known as the anthrax bacillus. However, his view was not accepted. There appeared to be instances of the disease where no anthrax bacilli were present. However, some brilliant experimental work first by Koch and then by Pasteur turned the tide in favour of Davaine's hypothesis. Koch showed in 1876 that the anthrax bacillus formed spores, which could survive in difficult circumstances to turn back into bacilli when conditions were favourable. This explained the existence of 'anthrax fields' in which grazing cattle nearly always caught the disease. In 1877, Pasteur showed experimentally that the apparent counter-examples to Davaine's hypothesis were actually instances of a disease other than anthrax. He, and his colleagues Chamberland and Roux, then set to work to produce a vaccine against the disease by attenuating the virulence of anthrax bacilli. All this evidence of the mechanism

of anthrax, moved the community to take a more favourable view of Davaine's hypothesis, but it only became completely accepted when some striking evidence of correlation was produced by Pasteur.

This evidence of correlation consisted of a randomised controlled trial conducted at Pouilly-le-Fort in 1881. 50 sheep were divided randomly into two groups. The first group were given the new vaccine against anthrax, while the second group were unprotected. All 50 sheep were then given a fatal injection of anthrax bacilli. The 25 unvaccinated sheep all died of anthrax. 24 out of 25 vaccinated sheep were perfectly healthy. Only 1 vaccinated sheep was sickly and later died, but it turned out that this sheep was a pregnant ewe, which died from complications of the pregnancy rather than anthrax. (It is worth noting that evidence about the mechanism of anthrax was needed for Pasteur and his colleagues to prepare the vaccine and to set up the randomised controlled trial. As discussed in §2.3, evidence of mechanisms is often needed to set up and interpret an RCT.)

These two case studies show how evidence of correlation and evidence of mechanisms can each be insufficient to establish a causal claim on its own—the claim only becomes established when the two sorts of evidence are both present. Detailed analyses of more recent cases of causal discovery also support RWT. Case studies include: establishing that the Epstein-Barr virus is a cause of Burkitt's Lymphoma (Clarke, 2011a); establishing that the human papillomavirus is a cause of cervical cancer (Clarke, 2011a); establishing that smoking causes lung cancer (Gillies, 2011); failing to establish that heavy drinking causes lung cancer (Gillies, 2011); establishing smoking as a cause of heart disease (Gillies, 2011). Russo and Williamson (2012) argue that the methods of the FP7 EnviroGenomarkers project (2009-2013), which aims to find biomarkers for environmental causes of disease, also fit with RWT.

Finally, surveys of present-day research papers have been put forward to support RWT. Russo and Williamson (2007) make a case for the thesis across the health sciences; Russo and Williamson (2011b) argue for the thesis by appealing to the practice of autopsy; Darby and Williamson (2011) cite papers in biomedical imaging as evidence for the thesis.

¶ *Uses of causality.* A consideration of the uses of causal claims allows one to see why the epistemological thesis RWT might be true. Causal claims are used in characteristic ways for prediction, explanation and control. Their use for prediction and control requires correlation: in order to predict an effect from a cause or vice versa, or to control an effect by controlling its causes, there needs to be a correlation between cause and effect, for otherwise neither variable on its own could provide any information about—or make any difference to—the other. On the other hand, the use of causal claims for explanation requires a mechanistic connection: arguably, the best way to explain an instance of B is to point to a mechanism showing how its causes are responsible for its occurrence; causal explanations are only explanatory to the extent that they can be viewed as providing a glimpse of the structure of a corresponding mechanistic explanation (Machamer et al., 2000; Williamson, 2013). Given the ways in which causal claims are used, it is thus no mystery that one normally needs to establish both correlation and mechanistic connection.

¶ Clearly, if RWT is true then there is good reason to treat mechanistic evidence alongside evidence of correlation when trying to test causal claims in medicine. This provides grounds to revise current advice provided by hierarchies of evidence,

which regard mechanistic evidence as inferior to—even trumped by—evidence of correlation. While we think that there is overwhelming support for RWT, it must be noted that RWT goes against some recent trends in both the philosophy of causality and in the methodology of medicine and it is controversial (see, e.g., Weber, 2009; Broadbent, 2011; Campaner, 2011; Dragulinescu, 2011; Howick, 2011; Campaner and Galavotti, 2012; Claveau, 2012). Hence it is only one of the grounds we cite for taking evidence of mechanisms more seriously. We shall now consider some other reasons.

§2.3. The design and interpretation of statistical trials

Despite the shortcomings of statistical trials that we have identified above, these methods often remain essential for gathering evidence to guide clinical practice. How useful this evidence is for guiding the care of the individual patient depends (amongst many others factors) on the design of the clinical trial(s) in question, and the way(s) in which data arising from trials are interpreted. The next subsection deals with the first of these questions, arguing that evidence of mechanisms is often required in order to design useful clinical trials. Evidence of mechanisms is also often required in order to interpret the data produced by a trial—a point which will be discussed in §2.5 and §3.1.

§2.3.1. Producing useful evidence from trials depends on diagnosis

> Physiological knowledge is not only indispensible in explaining disease, but is also necessary to good clinical observation. For example, I have seen observers surprised into describing as accidents certain thermal phenomena which occasionally result from nerve lesions; if they had been physiologists, they would have known how to evaluate morbid symptoms which are really nothing but physiological phenomena. (Bernard, 1856, p.200).

Statistical trials test for correlation in a specific population, which is usually defined in terms of particular diagnostic criteria. If a trial is designed with the intention of evaluating the cardiovascular effects of a novel treatment for high blood pressure, for instance, then these diagnostic criteria would be expected to specify appropriate clinical conditions for trial participation (having high blood pressure of a particular magnitude, for example). The results of this trial are then the correlations that obtain between treatment and clinical outcomes in this specified population. It will be clear that these correlations depend on the way in which the diagnostic criteria, used to collect the sample population, are specified. In the case above, changing the diagnostic criteria to collect a new sample population with extremely high blood pressure will lead to very different trial outcomes from those in the case where trial entry criteria recruit those with only moderately elevated blood pressure. In a more visible way, a closely-related process of diagnostic specification is used to control trial populations for the existence of confounding factors. Clearly our hypertension trial above will have very different outcomes in the pair of cases where the trial population contains, or does not contain, individuals with diabetes. In short, the results of clinical trials depend on diagnostic methods.

While purely clinical features are used to do this diagnostic work in clinical trials, as the quotation above suggests, evidence of mechanisms unsurprisingly plays a role in most. One good example is the changing way in which several cancers—most notably breast cancer and malignant melanoma—have been diagnosed and classified in

recent years. This change has been driven by the discovery of various causal genetic abnormalities in both these tumours. As the set of mutations which a particular tumour possesses is of major therapeutic and prognostic significance, classifications of these diseases increasingly feature some consideration of these genetic factors Clarke (2011b). In turn, this kind of reclassification, based on evidence of mechanisms, feeds back into the design of clinical trials, as our example of streptomycin (in §3.1) shows.

§2.4. External validity

§2.4.1. The problem of external validity

Here, we argue that evidence of mechanisms helps with the problem of external validity too. First, we recall the main lines of the debate on external validity in social science methodology and in philosophy of science. We then turn to two types of challenge: exporting treatments and exporting policy actions. We argue that in either case evidence of mechanisms is of great help in establishing whether and to what extent causal claims are valid externally.

'External validity', or extrapolation refers to the problem of exporting the results or methods of one study to a different population or setting. This a problem well-known and thoroughly examined in social science methodology since the Sixties and Seventies. Philosophers have only recently paid attention to the issue and formulated an alternative position.

Simply put, the original, methodological literature sees external validity basically as a problem to be resolved within statistics. In fact, the main threats to external validity, according to Cook and Campbell (1979) lies in the representativeness of the sample and in the possibility of replicating the study. The philosophical literature, instead, tried to go beyond this view and draw attention to the role *mechanisms* play in extrapolation. In particular, we owe to Steel (2008) the idea that successful external validity inferences can be made if we compare the mechanisms acting in the observed population and in the target population, especially at the most 'critical' points, that is where the mechanisms in the observed and in the target population are most likely to differ.

The problems described by Cook and Campbell, and subsequently by Steel, apply to RCTs too. In fact, even if we grant the soundness of an RCT, the question remains about its applicability *outside* the population of reference. The statistical 'set up' of an RCT is such that it maximises internal validity—namely if all goes well, we can establish with confidence that a treatment is effective in the population under examination. However, there is no a priori reason why the results of an RCT should be straightforwardly applicable to *another* population. The problem of 'exporting' the results of an RCT concerns both medical treatments and policy actions, as we shall now see.

§2.4.2. The external validity of treatments

Let us consider 'exporting medical treatments' first. The limitations of RCTs (with respect to external validity) are thoroughly discussed by Victora et al. (2004). The authors point to several issues that hinder the external validity of RCTs. In particular, they dispute that the internal validity of an RCT also ensures its generalisability. To be sure, this kind of argument is often invoked in the literature: the more a model is internally valid, the less it will be externally valid, and vice-versa. In social science, this argument has intuitive appeal because models that are tailored to the

background and the measurements of a specific population are of course too specific to say something sensible about *other* populations.

In RCTs, the assumption that results will be widely applicable—even outside the population of reference—follows, Victora et al. (2004) explain, from the assumption of 'universal biological response', i.e. different individuals will respond to a treatment or drug in the same way. The authors argue that although this assumption might well hold for "interventions with short causal pathways", it is certainly not the case for "interventions involving long, complex causal pathways, or in large-scale evaluations where these pathways can be affected by numerous characteristics of the population, health system, or environment", such as policy interventions. In fact, there might be two threats to successful extrapolation in the case of policy: one is "behavioural effect modification" and the other is "biological effect modification" (i.e., respectively, "differences in the actual dose of the intervention delivered to the target population" and "differences in the dose-response relationship between the intervention and the impact indicator").

Evidence of mechanisms helps to ascertain the external validity of treatments. Evidence of mechanisms can indicate how the intervention works (or is supposed to work) in the test population and whether, and to what extent, such mechanisms are also present in the target population (i.e., outside the trial)—see Steel (2008). A good example of this can be seen in clinical guidelines governing prescribing practices for antihypertensive drugs in the UK. Recent research has suggested that different drugs should be used for patients from different ethnic groups. NICE guidelines therefore state that treatment should differ, depending on ethnicity:

> Offer step 1 antihypertensive treatment with a calcium-channel blocker (CCB) to people aged over 55 years and to black people of African or Caribbean family origin of any age.... NICE (2011c, 5)

This recommendation was based on RCTs that had been designed to test the efficacy of different treatments in these ethnic groups. In turn, these trials were based upon the plentiful evidence suggesting the operation of different pro-hypertensive mechanisms operating in different ethnic groups—see NICE (2011b, 248-250 and citations).

One example of these trials was Kshirsagar et al. (2006). This involved 8960 study subjects, drawn from the general population, and followed up for a mean of 11.6 years for their risk of developing adverse CVD events. Among many other outcomes, this study showed that the risk of cardiovascular disease was higher for black people of African and Caribbean descent, and higher in older people (55–64 compared with 45–54). This evidence of different outcomes in different demographic groups was then used to support the clinical guidance quoted above, both as part of RCTs designed to test antihypertensive treatments, and as part of the *post hoc* analysis of evidence performed while writing the clinical guidelines.

§2.4.3. The external validity of policy actions

Let us now consider problems in 'exporting' policy actions.

¶ *Bangladesh nutrition.* Cartwright (2011) discusses the Bangladesh Integrated Nutrition Policy (BINP) as an example of unsuccessful exporting of policy actions. The point at stake is that while BINP largely failed to have an impact on child nutrition,

a very similar programme called TINP (the Indian Tamil Nadu Integration Project) in Tamil Nadu proved highly successful. How can similar policy actions have very different results?

Cartwright makes the point that policy makers neglected the different social structure of the populations to which they applied the programme, and this explains the success in one case and the failure in another case. TINP aimed to induce changes in mothers' behaviours in order to improve children's health status and nutrition. Such an intervention is based on understanding the social practices, norms, and habits—the social mechanism[4]—at work in that context. However, the same strategy did not prove successful in Bangladesh because in that context it is not the children's mother who takes care of the shopping and of preparing the meals, but the children's paternal grandmother. Consequently, targeting mothers turned out to be an inefficient way of improving children's nutrition. The social mechanisms at work in the two populations are different and consequently the same intervention does not achieve the same results.

Evidence of mechanisms helps establish the validity of policy actions because it adds precious knowledge about the similarities between the test and target populations, which is precisely the approach advocated by Steel (2008).

¶ *North Karelia.* The 'North Karelia Project' (Puska et al., 2009) illustrates the challenges of policy actions, but for a slightly different reason.

A massive public health action was launched in the 1970s in North Karelia (Finland) to reduce coronary heart disease mortality rates. Numerous activities were carried out, from putting health services in contact with individuals to massive media campaigns about healthy dietary habits. This policy action aimed to change health trends by changing the habits (consumption, diet, physical activity, etc) of individuals in the target population. In other words, it was the whole causal structure behind mortality rates due to coronary heart disease that was being altered. The results were good, but at the beginning, net changes due to the interventions could not be identified so clearly. The problem lay in the comparison with data coming from the neighbouring province of Kuopio, which was chosen as the 'control' area. Positive changes in coronary heart disease mortality rates were observed in both the target and the control area. The reason is that people in Kuopio were influenced by the programmes carried out in North Karelia. The intervention modified not only the causal structure in North Karelia (intended) but also in the province of Kuopio (unintended). While the public health action had correctly identified the mechanisms upon which to intervene, it was not anticipated that it would have *also* altered the causal structure of the control region.

In social science this is a well-known problem. Lucas (1976) was concerned with the validity of predictions about the effects of economic policies, on the basis of the 'known history'. The reason is that predictions are made using the 'known' structure, and since economic policy will change that structure, predictions will eventually be incorrect, once the policy is implemented. Steel (2008) calls this type of intervention 'structure-altering', precisely because they change the structure.

It is worth mentioning that the North Karelia Project raised important issues about exporting policy actions to different populations (see e.g. Wagner (1982); McLaren et al. (2006)). In particular, what is at stake is whether or not, in replicating

[4]There is an established tradition in sociology and social science on social mechanisms, see for instance the classic text Hedström and Swedberg (1988) and the recent contribution Demeulenaere (2011).

the action in different populations, their unique features are taken into account. This would lead to adjustment of the action to 'local' needs and habits, including different social practices, namely (social) mechanisms.

In sum, it is far from obvious whether a treatment will be efficacious outside the population in which it has been tested, or whether a successful policy action will be as good in a different context. No matter how well RCTs are designed and implemented, they do not on their own allow one to establish external validity. Evidence of mechanisms supplies information crucial to setting up the study and deciding how to adjust a policy action for a different population.

§2.5. The problem of inferring from the population to the single case

There is also another sense in which external validity poses a problem. Above, we discussed the inference from one population to another population. Here, the issue concerns the inference from the population (studied in the RCT) to a particular patient. While it is a merit of the evidence-based movement to have fostered protocols for treatment in order to ensure standardisation and comparability, there is no a priori guarantee that an individual patient will be similar enough to the average individual of the RCT to ensure that s/he will respond to the treatment in the same way. In such cases, considerations to do with single-case individual responses will be vital to support a claim that the same treatment will work in the single case. Thus one needs to know which mechanisms, or features of mechanisms, are instantiated in the particular patient. Again, statistical evidence works better when integrated with evidence of mechanisms.

§2.5.1. Precision, specificity and the reference class problem

One good example of the need to integrate evidence of mechanisms with evidence arising from statistical trials can be illustrated by taking account of the *reference class problem*:

> If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the problem of the reference class. (Reichenbach, 1949, 374)

As a toy example of the reference class problem, consider a calculation of the probability of an individual winning the Nobel prize for literature. Depending on the reference class to which we assign an individual, our individual probability will vary greatly. For example, if we assign our individual to the reference class of the human race as a whole, our individual probability will be very much lower than the probability we would infer if we assigned our individual to the class of successful authors. This, in turn, would be very much lower than the individual probability if our individual was assigned to the class of authors whose work had been previously considered by the Nobel prize committee.

The question of which reference class to choose is the source of the problem of the reference class. Various solutions have been identified. For example, Wesley Salmon suggested that we might resolve this difficulty by use of the *Reference class rule*, which he defined as follows:

> . . . choose the broadest homogeneous reference class to which the single event belongs. (Salmon et al., 1971, 43)

But this term "homogeneous" is a problem. How homogeneous is sufficiently homogeneous? Salmon's answer:

> To say that a reference class is homogeneous—objectively homogeneous for emphasis—means that there is no way, even in principle, to effect the relevant partition. (Salmon, 1977, 399)

This version of the reference class rule is clearly inadequate when it comes to clinical trials.[5] When designing clinical trials, it is not possible to adopt a highly granular design methodology that looks to test treatments in objectively homogenous reference classes. This is because of the vast numbers of individual descriptors that might plausibly affect outcomes. For instance, consider the number of different classes required to produce objectively homogeneous data for guiding the individual care of a 55 year old, left handed male, who worked as a painter and decorator in his twenties, but who is now currently unemployed, who smokes 10 cigarettes per day, with high blood pressure.

While this kind of problem is presented in the epidemiological literature as a question of precision against cost,[6] the more fundamental problem here is one of the number of subjects available: if we want this kind of specificity at any reasonable level of precision, we would need trials that involve multiple copies of the species as a whole. So the reference class rule as formulated above is of no practical use in these cases. Given that this objective formulation of the reference class rule is so demanding, Salmon provided a more measured alternative—the epistemic formulation of the reference class rule. This states that a reference class is epistemically homogeneous when we do not know how to make any statistically relevant partitions (Salmon et al., 1971, 44).

One means of achieving this partitioning is to use features suggested by a relevant mechanism. Given a reasonably well-confirmed mechanism, it is possible to characterise different subgroups in a trial by their mechanistically salient properties. For example, in the case of Kshirsagar et al. (2006), groups of subjects belonging to particular ethnic groups, or of particular ages, were analysed separately, on the grounds that evidence of mechanism provided good reasons to suspect that these groups would respond to the drug under test rather differently from each other. Here then, these groups are interpreted as if they were satisfying Salmon's epistemic formulation of the reference class rule (Clarke, 2011a, 177ff). We might adopt the following slogan for this kind of practice: determine your reference classes by looking to evidence of mechanisms.

[5] The remainder of this section will deal with the reference class problem as applied to trial design. However, as the epidemiological terminology differs significantly between trial design and interpretation of trial data, it's worth making a few brief remarks here regarding the problem as it affects interpretation, particularly the practice of stratification. Stratification involves dividing up trial results to examine outcomes in partitions of the trial population thought to be interestingly different from the general population. For instance, a trial of an antihypertensive agent might stratify the trial population into age groups at the data analysis stage, to see if the drug response differs. This kind of practice is limited by various kinds of information bias, including the *sparse data problem*: the smaller the strata, the greater the variability of apportionment ratios (Rothman et al., 2008), and the lower the precision of any resulting causal claims. In short, the problem is identical in either trial design or interpretation.

[6] And see the literature on sample size calculation for examples of this: Rothman et al. (2008, 149ff).

§2.6. The limitations of evidence of mechanisms

Evidence of mechanisms can be very helpful in all the ways outlined above. Here, however, we examine three qualifications to the use of evidence of mechanisms. First, stories about mechanisms can be overly psychologically compelling—it is *evidence* of a mechanism, rather than a story, that is required. Second, even good evidence of mechanisms might suggest a mechanism that is so complex it is hard to tell whether A will make a difference to B and if so in which direction (positive causation or prevention) and how much of a difference. Third, finding evidence of one mechanism does not rule out the existence of other mechanisms that mask its effect. We show how supplementing evidence of mechanisms with evidence of correlation is often exactly what is needed to mitigate the latter two problems.

§2.6.1. Psychologically compelling stories

The first major problem worth emphasizing is that *stories* about possible mechanisms of action can be psychologically compelling when they shouldn't be—psychologically compelling to patients, but also to experienced researchers and practitioners. Because of this, such stories have led to bad mistakes in the history of medicine. For example, bloodletting as a treatment for multiple illnesses was based on a story about the action of the human body that was quite wrong (Carter, 2012). Many commentators on evidence in healthcare, and specifically the promotors of the evidence-based movement, have been sceptical of the place of mechanistic evidence for precisely this reason—and they are absolutely right to urge caution.[7] Because of this, it is crucial that evidence of mechanisms should be treated explicitly, and not allowed to drive thinking implicitly. Two points address this problem. First, what is relevant to evidence in healthcare is not stories about mechanisms, but *evidence* of mechanisms. We examine what this involves in detail in section §4. Second, even excellent evidence of mechanisms is most effective when treated as complementary to evidence of correlation, rather than standing alone. This is particularly due to the two following problems: the complexity problem and the masking problem.

§2.6.2. The complexity problem

Even where a mechanism linking A to B is well established and known in some detail, it can be hard to infer whether A has a positive effect on B, or A prevents B, or indeed whether A has any net effect on B at all. This is particularly true in cases where the mechanism is complicated: where there are several links on a pathway from A to B or where there are several pathways from A to B. It is also a problem where a mechanism is known to be non-robust over time or over other changes in situation. It is typically evidence of correlation that is crucial for determining whether any causation is positive or negative and what the net effect is. Thus evidence of mechanisms should be used in conjunction with evidence of correlation, not on its own, to infer causal claims.

Multifactorial diseases provide a rich seam of examples that illustrate the difficulties of successfully working in these complex environments. One particularly fine case is that of dalcetrapib, which exploits the finding that individuals with high HDL:LDL ratios have much lower risks of developing coronary heart disease than

[7] Howick (2011) provides other examples of such compelling stories and how they led to the development of EBM.

individuals with low ones. The usual therapeutic emphasis has been to improve this ratio by reducing serum LDL, particularly by the use of lifestyle interventions and statins. Dalcetrapib promised a novel way of improving it by increasing HDL directly. Given the clinical importance of heart disease, and the substantial financial rewards for finding effective strategies for preventing it, the drug therefore received a great deal of attention in the medical (and financial) press during its development. Clinical testing of the drug was initially extremely promising. Trials showed, first, that it was effective in increasing serum HDL concentrations in study populations (Stein et al., 2010). Second, that administration of the drug was correlated with a small decrease in the extent of atherosclerotic plaques in research subjects (Fayad et al., 2011). However, hopes were dashed when a large study, known as dal-OUTCOMES (Schwartz et al., 2012), was terminated in May 2012 on the grounds of futility. Study investigators had failed to find any improvement in actual CVD outcomes in subjects taking dalcetrapib compared to placebo. While the reasons for this failure are still somewhat puzzling, it seems likely that some other pathways linking HDL and CVD are actually the effective determinants of cardiovascular outcomes, rather than the more direct link between high HDL and reduced CVD risk previously thought to be responsible. Far from being a major net contributor to the effect, this factor seems to contribute to the effect not at all.

§2.6.3. The masking problem

The masking problem provides a further reason why even very good evidence of mechanisms must be treated with care.

To see the problem, suppose you have very detailed evidence of a mechanism linking A and B. E.g., you have found the bacteria and understood how they cause the disease, you have studied a drug and found that it kills the right bacteria, and doesn't harm people, and you are confident that killing the bacteria will cause full recovery. So you can trace the mechanism all the way from the drug to recovery, or trace the process, in Steel's terms (Steel, 2008). However, you cannot conclude that taking the drug will cause recovery. This is because finding one mechanism linking A and B does not prove that there are *no other mechanisms operating*.

The human body is a complex system, and the more we discover about it the more it seems that it is very common to have multiple mechanisms operating. If there are multiple mechanisms operating, they may impact on each other, and one or more may *mask* the effects of the mechanism you have discovered. Steel uses the example of the relationship between exercise and weight loss (Steel, 2008, p. 68). We know well how increased exercise burns calories and leads to weight loss. But we also know that exercise increases appetite, and we know that eating more leads to weight gain. Unless we investigate further, we don't know which mechanism will, as it were, 'win'. So we don't know whether increasing exercise will, on average, make you fat or make you thin.

Given the deep entanglement of mechanisms in the body, we may be unable even to be sure of a 'component effect', which Hitchcock opposes to the 'net effect' described above (Hitchcock, 2001). Thus an analogue of the complexity problem that occurs within a single mechanism (§2.6.2) also occurs outside the mechanism, within a system of mechanisms. Happily, we already have well-developed ways of dealing with that sort of problem, because what is required to know what happens overall, on average, is evidence of correlation. Evidence of correlation tells you which mechanism 'wins', which one masks the other.

§2.6.4. What to do about these problems

What this means is that evidence of mechanisms and evidence of correlation do not act independently, each suggesting separately that A does cause B. They integrate in a special way. To summarize (updated from (Illari, 2011, p147.)):

1. Evidence of a correlation relation between A and B:

   Its problems are that of confounding and non-causal correlations.

   Its advantage is that it can reveal masking, and can help assess the net effect of a complex mechanism.

2. Evidence of a mechanism linking A and B:

   Its problem is masking, and being too complex to assess a net effect.

   Its advantage is that it can reveal confounding and non-causal correlations.

Evidence of a linking mechanism helps show that the overall relationship between A and B is genuinely causal. But evidence of correlation helps to determine the net effect of a mechanism, and to show that it is not masked by further unknown mechanisms. Together, evidence of these two different things is very much stronger than evidence of one alone.

We can describe this situation by an analogy to reinforced concrete, which is formed by placing steel grids into concrete. Now most concrete mixes have high resistance to compressive stresses, but any appreciable tension (e.g., due to bending) will break the microscopic rigid lattice, resulting in cracking and separation of the concrete. Steel, however, has high strength in tension. So, if steel is placed in concrete to produce reinforced concrete, we get a composite material where the concrete resists the compression and the steel resists the tension. The combination of two different materials produces a material that is much stronger than either of its components. In the same way, we argue that it is the combination of two different types of evidence which produces much stronger overall confirmation than would either type of evidence on its own. The important point is that this depends on the evidence being evidence of two types of things—correlations and mechanisms—that are different in character. This is why when, in section 4, we try to integrate evidence of mechanisms and evidence of correlation, we do not do so in terms of a single hierarchy but in terms of two interconnected grading procedures (see Figure 5 below).

§3
Integrating evidence of mechanisms and evidence of correlation in practice

So far we have divided evidence into evidence of correlation, such as is obtained from RCTs, observational studies and so on, and evidence of mechanisms, which is often obtained from laboratory experiments. We have argued that to evaluate a causal claim in medicine, evidence of mechanisms should be considered alongside evidence of correlation. In this section, we will show in some detail that there are often severe problems with trying to rely on RCTs alone without taking account of evidence of mechanisms. We will focus on illustrating the point made in §2.3 that, in many cases, evidence of mechanisms is needed both to design a RCT, and to interpret the results which it gives.

§3.1. Streptomycin

We will argue for this claim by considering in detail a particular example, namely the two trials of streptomycin and other anti-tuberculosis chemical agents, which were carried out by the British Medical Research Council (MRC) in the period 1947-50.[8] These trials are of considerable importance in the history of medicine, because they were among the first RCTs, and they were one of the strong influences which led to the increasing use of RCTs to test the efficacy of proposed medicines.

Streptomycin was discovered in America in 1944 by Schatz, Bugie, and Waksman. It was shown that it strongly inhibited tubercle bacilli *in vitro*, and that it was also successful *in vivo* in treating experimental tuberculous infections in guinea-pigs. The new antibiotic even produced some quite spectacular cures of patients suffering from tuberculosis. In fact so promising did streptomycin appear that it might have seemed immoral to conduct a randomised controlled trial of the new antibiotic, since those who were unlucky enough in the random allocation to be assigned the then standard treatment (prolonged bed-rest) might thereby have been deprived of an excellent hope of cure. Possibly for this reason, no controlled trial of streptomycin in pulmonary tuberculosis was undertaken in 1946 in the U.S.A. In England, however, an influential medical statistician (Austin Bradford Hill) was a firm believer in the necessity of randomised control trials, and managed to persuade the Medical Research Council (or MRC) to carry out an RCT. The first patients for it were recruited in January 1947.

The report on the trial published in the British Medical Journal on 30 October 1948 (MRC, 1948) contains an account of both the procedure and the results. The procedure was fairly straightforward. The first requirement was to make the patients and their disease as uniform as possible. The type of case to be considered was therefore defined quite precisely as follows (MRC, 1948, p. 770):

> acute progressive bilateral pulmonary tuberculosis of presumably recent origin, bacteriologically proved, unsuitable for collapse therapy, age group 15 to 25 (later extended to 30).

Between January 1947 and September 1947, 109 patients had been accepted. 2 of these died in the preliminary observation week, and the remaining 107 were assigned randomly to either the control group C or the streptomycin group S. There were 52 in C, and 55 in S. The control group C received the standard treatment of the time, which was 6 months of bed-rest. The S group received, in addition to bed-rest, a dose of 2g of streptomycin per day, given in four injections at six-hourly intervals. The streptomycin was continued for four months, but the patients were observed for a further 2 months. So the trial was brought to a close for each patient after 6 months.

The improvement or deterioration of the patients in the 6 months of treatment was assessed by X-rays, 'the radiological picture … being in our opinion the most important single factor to consider' (MRC, 1948, p. 771). The results obtained are shown in the following table (MRC, 1948, p. 771).

[8] MRC (1948, 1949, 1950) contain the reports on these trials published in the British Medical Journal. There is an overview of the trials in Daniels and Bradford Hill (1952), while Bradford Hill (1990) gives some interesting reminiscences.

Radiological Assessment	Streptomycin Group	Control Group
Considerable improvement	28 (51%)	4 (8%)
Moderate or slight improvement	10 (18%)	13 (25%)
No material change	2 (4%)	3 (6%)
Moderate or slight deterioration	5 (9%)	12 (23%)
Considerable deterioration	6 (11%)	6 (11%)
Deaths	4 (7%)	14 (27%)
Total	55 (100%)	52 (100%)

These figures show that the S group did very considerably better than the C group. 51% of the S group showed considerable improvement as against only 8% of the C group. 7% of the S group died as against 27% of the C group. These differences are highly significant statistically.

In the light of such good results from the RCT, one might have expected that the MRC would have declared that treatment with streptomycin had been shown to work, and was to be recommended. Instead of giving such an endorsement of streptomycin therapy, however, the MRC conclude on a very cautious note, saying (MRC, 1948, p. 780):

> This planned group investigation has demonstrated both the benefit and the limitations of streptomycin therapy in pulmonary tuberculosis.

This caution proved to be amply justified. The same patients were investigated after 5 years, and it was then found that 58% of the S group had died as against 67% of the C group. The difference here is not statistically significant. These figures are taken from Florey (1961, p. 133), where she comments: 'it was obvious that the encouraging promise at an earlier time had not been fulfilled'. It should be noted that the patients were all between 15 and 30 years old at the start of the trial. So it is unlikely that any of them would have died from causes other than tuberculosis in the succeeding 5 years. What seems to have happened in the S group is that, after the encouraging initial improvement, many relapsed.

This example shows that there is a general problem with RCTs. Such trials have to come to an end after some time period t. Suppose the RCT shows that the treatment has produced a marked improvement by t, can we then be sure that this will not be followed by a relapse later on?

How can this problem be overcome? Well, those who conducted the streptomycin trial did seem to overcome the problem. They did foresee that the long-term results might not be so good as was suggested by the short-term improvements; and, for this reason sounded a note of caution. How did they manage this? The answer is that they took account of evidence about the mechanism of the treatment. This is perhaps not surprising because of the involvement of Bradford Hill in the trial. As we pointed out earlier, Bradford Hill's criteria for establishing causal relationships in medicine included both evidence of correlation and evidence of mechanisms. The importance of mechanisms was not likely to be forgotten by any group of which he was part. Here now is the analysis which was made of the mechanism of the streptomycin treatment.

Already by 1947 many researchers in the area had become aware that there might be a problem with streptomycin therapy (see Florey, 1961, pp. 136–7). While some antibiotics such as penicillin could dispose of the pathogenic bacteria, which they targeted, in a week or two, streptomycin took many weeks, even months, to deal with a patient's tubercle bacilli. Now Darwinian evolution as applied to bacteriology

strongly suggested that, in such a time period, strains of the tubercle bacillus might develop which would be resistant to streptomycin. Such resistant strains posed a very considerable threat to streptomycin therapy. They might well increase in numbers producing a relapse, and, in this new condition, a fresh treatment with streptomycin would obviously be useless.

Because of an awareness of this difficulty, those who carried out the streptomycin RCT, at the same time carried out an investigation into the mechanism of the treatment. They tested the resistance of the tubercle bacilli in patients who had been given streptomycin. It emerged that, by the end of the second month, 63% of the cases in the S group, which were examined, had developed resistance to streptomycin.

The thinking of researchers in the MRC group in the light of this evidence about the mechanism of streptomycin treatment is clearly stated in MRC (1949, p. 1521):

> A major disadvantage in the use of streptomycin is that the period of effective therapy is limited in many patients by the emergence of streptomycin-resistant strains of tubercle bacilli after five or more weeks of treatment. It has been thought by many workers that the addition of another tuberculostatic agent might be sufficient to suppress the resistant strains, which in the initial phases are present in very small numbers.

In fact Jorgen Lehmann, a Danish doctor working in Sweden, had announced in 1946, the existence of a tuberculostatic agent, namely para-amino-salicylic acid or PAS.

As soon as the first streptomycin trial was over, the MRC researchers started a second trial. This was conducted along similar lines to the first trial, except that the patients were divided into 3 groups. The S group received streptomycin only, though only 1g a day, given by one injection at 8am. The P group received 20g daily of PAS by mouth in four doses of 5g at 8am, noon, 4pm and 8pm. The SP group received both streptomycin and PAS in doses as for the other two groups. These treatments were continued for 3 months, and the patients were observed for a further 3 months.

Patients improved most in the SP group, but the most striking results concerned the difference in streptomycin resistance between the S group and the SP group. As it is stated in MRC (1950, p. 1081):

> At the end of the six months 89% of the SP patients producing positive cultures had completely sensitive strains, and only 21% of the S patients.

Moreover the resistant strains took a longer time to appear in the SP Group, and some subsequently disappeared.

By December 1949, the results concerning streptomycin resistance were already so striking that an interim communication was made (MRC, 1949). The authors conclude their report on the full trial by saying (MRC, 1950, p. 1085):

> Combination of PAS with streptomycin not only renders effective administration of streptomycin possible for longer periods than previously, but probably permits also of repeated effective courses.

This concludes our account of the MRC RCTs, and we will now consider what conclusions can be drawn.

The first streptomycin trial showed that the patients given streptomycin improved dramatically over six months when compared to the controls. If this had been

accepted, and all evidence, concerning the mechanism by which the streptomycin therapy worked, had been ignored, then the conclusion would inevitably have been reached that streptomycin on its own was an excellent therapy for tuberculosis. However, the reality was that streptomycin on its own was, over a longer period, no better than bed-rest.

Fortunately, however, the MRC researchers of that period took it as a matter of course that both the results of the RCT in terms of patient improvement, and the evidence about the mechanism of streptomycin therapy should be taken into account, and, when both were weighed they reached the correct conclusion that there were problems with using streptomycin on its own as a treatment for tuberculosis. This is a very good instance of the epistemological thesis RWT, which, as we saw in §§2.2, states that a causal claim in medicine should in general only be accepted if it is supported both by evidence of correlation, *and* by evidence of mechanisms.

Considering both types of evidence not only enabled a serious error to be avoided, but it also suggested a way out of the problem which had been brought to light. This was the conjecture that the combination of streptomycin with PAS might prevent the emergence of strains of tubercle bacilli resistant to streptomycin. This conjecture was tested out, and proved to be correct. Together with further research and development, it led to satisfactory treatments for tuberculosis.

This example shows that evidence of mechanisms is needed both in the design of some RCTs in order to reach a decision about the length of time for which the trial should run, and also in the assessment of the results of the RCT.

§3.2. More recent developments

This section aims to develop in further detail the account of the integration of evidence sketched out in the historical case above. A defender of the idea that the statistical evidence produced by RCTs can be understood in isolation—something we might term the *naked statistics* view of EBM—might point to modern clinical trials of TB chemotherapy as uncomplicated examples of evidence of correlation alone providing sufficient evidence to ground meaningful decision regarding the care of individual patients. As we shall see, though, the naked statistics view is undermined by research that sought to determine what the most appropriate duration of chemotherapy for tuberculosis should be.

As the historical example above shows, this question of treatment duration is not a recent therapeutic concern. Once the problem of drug resistance had been recognised, various new antitubercular agents were introduced. These were typically used in regimens consisting of multiple antibiotic agents (Fox et al., 1999), on the grounds that combination therapy in general appeared to show greater treatment efficacy when compared with monotherapy. An important goal of this research studied treatment duration: how long should overall treatment last, and how long (and in what sequence) should each treatment be given for?

It was soon recognised that simply preferring longer treatment programmes to shorter ones was an unwise strategy. While increasing the length of treatment did seem to reduce the rate of tuberculosis relapse, *ceteris paribus*, this was counterproductive in terms of overall outcomes. Just to illustrate the complexity of the problem, the longer the treatment, the greater the expense of treatment and the greater the risk of the recipient of therapy experiencing adverse drug effects. Very lengthy treatment also significantly lowered the degree of patient concordance (reviewed, as part of a much more sophisticated analysis by Munro et al. (2007)), leading in turn to a

theoretically increased risk of drug-resistant strains of tuberculosis evolving in the population.

Difficulties in finding suitable surrogate markers capable of indicating when the disease was eradicated, when coupled with the delay between the end of treatment and the detectable resurgence of un-cured disease, meant that treatment durations initially erred on the long side. Most TB eradication regimens in the 1960s and early 1970s, lasted for between 12-18 months. However, research in the 1970s (Fox et al., 1999) led to the adoption of shorter, cheaper, safer, therapeutic strategies. These shorter strategies are now recommended as standard. For example, as the recent National Institute for Health and Clincial Excellence Clinical Guideline on TB (NICE, 2011a) notes, evidence supports the following treatment of respiratory TB:

> Six months of daily treatment with rifampicin and isoniazid, supplemented in the initial two months with pyrazinamide and either ethambutol or streptomycin (the six-month four-drug regimen) has been the evidence-based gold standard for TB treatment for at least the last 15 years. No new first-line drugs have been found for over 30 years. Attempts have been made to shorten the total duration of treatment by reducing the duration of the continuation phase of treatment. The comparators for such studies are the results of the six-month, short-course, four-drug regimen, which give a cure and completion rate of >95% and a relapse rate of 0-3% in both clinical trial and routine clinic use. Such controlled studies have been largely conducted in adults not known to be HIV positive, with a few in HIV-positive individuals or in children. (NICE, 2011a)

In sum, determining treatment duration remains an important role for evidence of mechanisms.

§4
Guidelines for integrating evidence of mechanisms and correlation

We have have argued in §2 that there is a need to treat evidence of mechanisms alongside evidence of correlation, as each addresses the major weakness of the other. We have built on that in §3 by seeing how evidence of mechanisms has had an important role in establishing and refining causal claims about the treatment of tuberculosis. In this section we shall offer some general suggestions for how evidence of mechanisms might be considered alongside evidence of correlation.[9]

§4.1. Evaluating all the evidence

Figure 5 provides a graphical representation of the points that were made in §2. The main claim is that evidence of mechanisms should be treated alongside evidence of correlation, rather than as inferior to it. To establish a causal claim, one normally needs to establish a mechanistic claim as well as a correlation claim (§2.2); hence one needs to grade the evidence in favour of the mechanistic claim, just as one

[9] In a sense, we follow up the suggestion of Solomon (2011) that EBM largely ignores basic science, particularly mechanisms, and we offer an account of how to integrate grading evidence of mechanisms and grading evidence of correlation.

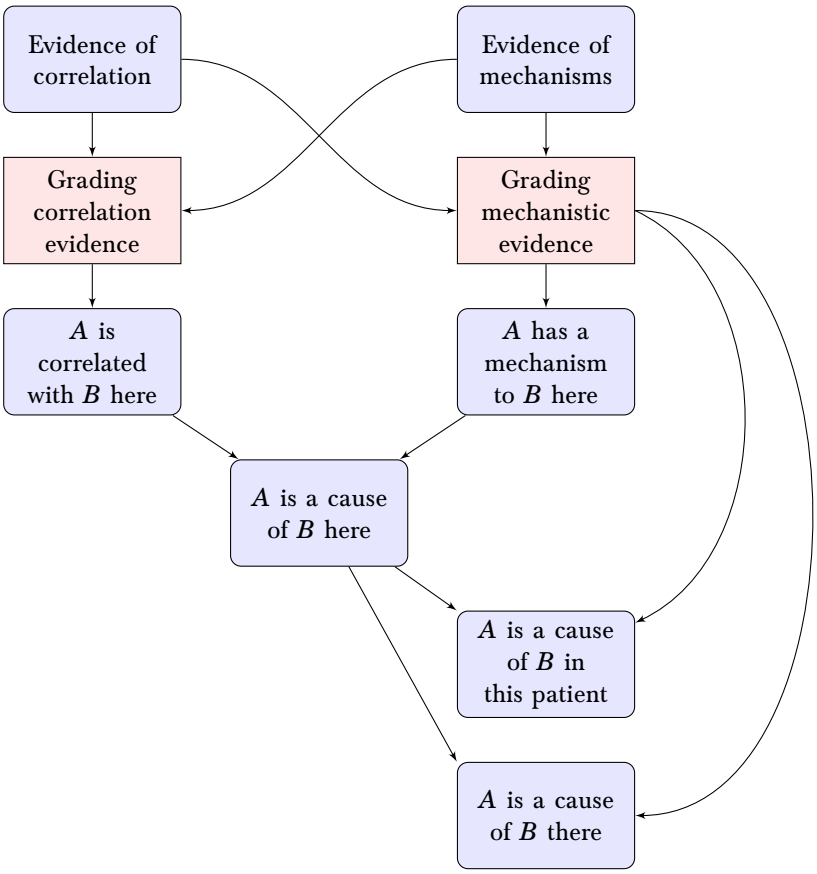

Figure 5: Evidence of mechanisms treated alongside evidence of correlation.

needs to grade the evidence in favour of the correlation claim (the square boxes in Figure 5). Evidence of mechanisms is needed to assess whether statistical trials have been properly designed and interpreted (§2.3). Thus there is an arrow from 'Evidence of mechanisms' to 'Grading correlation evidence' in Figure 5. On the other hand, evidence of correlation is needed to determine the net effect of a mechanism (§2.6.2), and to assess whether the mechanism is being masked by other mechanisms (§2.6.3). Thus there is an arrow from 'Evidence of correlation' to 'Grading mechanistic evidence' in Figure 5. Mechanistic evidence is also important in applying a causal claim to a single case (§2.5), or to another population (§2.4), as represented by the curved arrows on the right-hand side of Figure 5. Note that each of these two sorts of application will require its own protocol for grading the relevant evidence of mechanisms (this distinction between protocols for grading mechanistic evidence is not represented in Figure 5).

The question remains as to how evidence of mechanisms can best be categorised and graded.

§4.2. What evidence of mechanisms is evidence of

We have seen that evidence of the entities, activities and organization of the mechanisms by which medical treatments and policy interventions work can be gained in many ways, such as direct observation, technologically assisted observation, simple manipulation, repeated experimental manipulation, simulation. (For further discussion see Darden (2006); Craver (2007); Bechtel (2008); and Bell (2008) for discussion of Watson and Crick's use of chemical mutagens to crack the DNA code.)

Consider some of our examples. In the work on cholera by Koch discussed above, evidence of the mechanism was obtained by examining levels of cholera vibrio in filtered and unfiltered water, while in the anthrax case, the anthrax bacillus was studied and its ability to form spores discovered. We have also seen how evidence of mechanisms and of correlation is integrated in practice in detail, in treating tuberculosis. There are numerous places where mechanistic evidence was useful there. General background knowledge of mechanisms was important. It was known that killing the bacteria causing a disease could cause cure. It was also known that some bacteria can become resistant to particular antibiotics, so that a sensitive strain can develop into a resistant strain if given enough time. But this is not yet evidence of what is happening in a particular case. What was of crucial concern to the streptomycin trials were tests—*in vitro*—of when the tubercle bacilli, the bacteria that causes tuberculosis, were being effectively killed by streptomycin, and when they weren't, due to developing streptomycin resistance. It is relatively easy to observe *in vitro* whether and how quickly a specimen of the bacilli gathered from a patient is killed by the treatment antibiotic. Even the measure of successful treatment used—x-rays of the lungs—involved seeking evidence of the mechanism of cure working. These kinds of investigations yield far more than a story about how an antibiotic works. They yield direct evidence that the hoped-for mechanism for recovery is in fact working in one particular patient.

While there is not the space here to do justice to the full diversity of sources of evidence of mechanisms, some examples are listed in Table 1 below.

Table 1: Examples of sources of evidence of mechanisms

Direct manipulation: e.g., in *vitro* experiments
Direct observation: e.g., biomedical imaging, autopsy, case reports
Statistical trials: e.g., RCTs
Confirmed theory
Analogy: e.g., animal experiments
Simulation: e.g., agent-based models

We shall now present one important way of categorising evidence of mechanisms linking a putative cause A with a putative effect B. We will also see how evidence of other mechanisms in the domain—mechanisms that do not themselves link A and B—might nevertheless bear on the question of whether A causes B.

We suggest there are three things one can get evidence of:

1. Evidence of the details of a *specific* mechanism linking A and B.

2. Evidence that there exists *some* mechanism linking A and B.

3. Evidence that there is *no* mechanism linking A and B.

The first case is the clearest case of evidence of mechanism being useful, and we have seen this in the tuberculosis case above. As is not uncommon, evidence of the correlation between treatment and cure was gathered in the same study as evidence of the mechanism for cure. This was the case in the original streptomycin trial, where tubercle bacilli samples taken from patients were tested for streptomycin resistance, indicating that the treatment would begin to be ineffective. The later trial of the combined treatment—SP—continued testing samples from patients, showing that many fewer patients on SP than on streptomycin alone had streptomycin-resistant bacteria. These trials were essential to discover whether the expected mechanism of recovery was acting, and how quickly, and enough knowledge of the mechanism was gained to allow decisions to be made about required length of treatment, and likelihood of complete cure.

The BINP case illustrates case 2, where there was initially reason to believe that there is a mechanism of the postulated kind in the Bangladesh population. The TINP case provided evidence that such a mechanism exists in Tamil Nadu, which in turn provided (rather weak) evidence that the analogous mechanism exists in Bangladesh. This case demonstrates that the strength of evidence of mechanisms matters: such evidence can let us down, when salient differences between populations go unnoticed. When we consider the social structure of the Bangladesh population, we might postulate that educating paternal grandmothers about nutrition will improve child nutrition. The combination of the effectiveness of the policy of educating those responsible for feeding the children in Tamil Nadu and the direct evidence of the different social structure of the Bangladesh population gives us good reason to believe this. It could be tested by running a trial educating Bangladeshi maternal grandmothers, *and* observing whether their behaviour actually changes during the trial. After these tests, if child nutrition improved, and the behaviour of maternal grandmothers had changed, we would have good evidence of the linking mechanism, as in case 1.

Case 3, evidence that there is no mechanism, is also important, as it is used to delimit the space of possible mechanisms of action, and so of possible causal relations between A and B. The retroactive intercessionary prayer case discussed in §2.2 is a nice example (Leibovici, 2001). Even many religious people do not believe that prayer is an effective cure when a loved-one has a serious illness. Very few would believe that praying for ill loved ones retroactively—after their recovery or death—would have any effect whatsoever. We have now found a great deal of evidence of many different kinds of mechanisms of disease causation, and cure, and the fact that current science renders the existence of a mechanism for retroactive prayer very improbable tells against the effectiveness of retroactive prayer.

§4.3. Grading evidence of mechanisms

In the paper up to this point, we have made the case that evidence of mechanisms can usefully supplement evidence of correlation. Of course, all evidence and all conclusions reached in medicine are fallible. Evidence of correlation is fallible; evidence of mechanism is fallible; and conclusions drawn from that evidence are fallible. That is the nature of any science. We focus here on the important point that we can get varying quality of evidence of mechanisms, just as we can get varying quality of evidence of correlation. We have been pressing the point that this kind of

variation in quality of evidence of mechanisms needs a great deal more attention—indeed, it needs just as much attention as quality of evidence of correlation. Here we make a very preliminary attempt to lay out some ways in which evidence of mechanisms may be graded. We acknowledge that much more work will need to be done in this regard.

We begin by illustrating in Table 2 below how evidence of a mechanism might be graded by looking for positive and negative aspects of evidence of the features of mechanism. Features of a mechanism include its entities, the activities or interactions of the entities, and the organisation of the mechanism. The table is not ordered, in that it is not meant to imply that indicators earlier in the table are more important. This is impossible, as different indicators will be more or less important in different contexts. The table is almost certainly incomplete.

Table 2: Grading evidence of a mechanism

Pluses	**Minuses**
Each independent method that confirms a feature	Each independent method that fails to confirm—or, worse, disconfirms—a feature
Each independent research group that confirms a feature	Each independent research group that fails to confirm—or, worse, disconfirms—a feature
Larger proportion of features found	Smaller proportion of features found
Analogous mechanisms known	The analogy is a weak one, or, worse, analogous situations exhibit no such mechanism
Robust, reproducible across a wide range of conditions	Fragile, not reproducible in slightly varying conditions

Such a table needs to be interpreted with a certain amount of common sense. The recommendations about multiple methods for detection of features of the mechanism has to be applied in the light of awareness of whether alternative techniques are in fact available. If they are, and they are unused, that may indicate a problem. It is worth investigating whether there is a sensible reason for a technique not having been applied. If no other techniques are as yet available, that is a very different situation. Similarly, it is not suspicious that a single research group confirms the result if that work has just been made public. It might, of course, still be reason for *caution*. However, as time goes on, if other research groups attempt to replicate the results, and fail, then naturally that tends to indicate a problem.

The degree to which an analogous mechanism confirms a mechanistic claim depends of course on the degree to which the two mechanisms are similar and the degree to which the two contexts in which the mechanisms are situated are similar. As we mentioned above, Bradford Hill recognised the importance of analogy in medicine in his ninth criterion (Bradford Hill, 1965). We should not overlook analogical evidence, because mechanisms postulated in this way can be vitally important for the development of medical knowledge. This is most obvious when an entirely new mechanism of cure or disease causation has been found. Before the discovery

that penicillin cures disease by killing the bacteria causing the disease, there would have been no reason to look for other agents that kill bacteria—agents such as streptomycin. Penicillin opened up the possibility of many other analogous mechanisms of cure. New mechanisms of disease are equally important. For example, reasoning by analogy was why the concern was raised about the length of time required for treatment allowing bacteria to develop resistance in the streptomycin case (see below). Further testing revealed that this concern was quite correct. Postulating and then finding analogous mechanisms of both cure and disease causation is thus an important way of increasing causal knowledge.

Moving on to fragility, note that some mechanisms just are fragile. It will be harder to use such a mechanism for causal inference, but this need not indicate that the mechanism is a mere artefact. For example, the mechanism of action of Dalcetrapib, discussed in §2.6.2, proved not to be robust enough to lead to reduction in coronary heart disease. This weakness does not mean that the mechanism detected *in vitro* was spurious, just that it did not ultimately produce the same effect *in vivo*.

Notice that good evidence of mechanisms does not have to be produced by a statistical trial. Single-case observations of a mechanism that are confirmed by several independent sources and methods can be excellent evidence of a mechanism. This means that assessing quality of evidence of mechanisms and quality of evidence of correlation are distinct tasks.

After these general remarks about the criteria of Table 2, it will be helpful to consider how they might be applied to a specific example. In our earlier discussion of the MRC trials of streptomycin (§3.1), we pointed out that the researchers found evidence of a mechanism operating in the treatment, namely the development of strains of tubercle bacilli resistant to streptomycin. Let us now see how the criteria of Table 2 might apply to the evidence for this treatment mechanism. First of all let us consider whether independent methods confirmed the feature (development of resistance). In the course of the first trial, the resistance of the tubercle bacilli was measured in a direct fashion by what is known as the resistance ratio (R.R.). This is the ratio of the minimum concentration of streptomycin to which the tubercle bacilli of the patient are sensitive to the corresponding figure for the standard strain H37Rv. In the second trial a combination of streptomycin with another bacteriostatic agent (PAS) was used. Since PAS would still be effective against the strains resistant to streptomycin, and since the combination would deal with the tubercle bacilli more quickly, allowing less time for resistance to develop, the prediction was that fewer resistant strains would develop in this version of the treatment. This proved to be the case providing an independent way of showing the role of resistance in the treatment mechanism. The second criterion concerns independent research groups, and so was not applicable immediately after the MRC trials, though, of course, the results concerning the development of resistance were confirmed later by other research groups. The third criterion concerns a larger or smaller proportion of the features postulated, but since we have only one feature here (development of resistance), this criterion is not applicable. The fourth criterion is that analogous mechanisms should be known. This provides strong evidence for the postulated treatment mechanism, since, in the light of Darwinian evolution, we would expect resistant strains to develop, and many similar Darwinian phenomena were known. The last criterion is that the mechanism should be robust, reproducible across a wide range of conditions. The development of resistance, to a greater or lesser extent, occurred in all the trials carried out, showing it to be robust phenomenon.

That concludes our account of the criteria of Table 2, but, finally, we should re-

iterate that evidence of mechanisms, graded for quality, is to be allied with evidence of correlation, similarly graded for quality, in evaluating a causal claim. Ultimately, the question is whether there is good reason to believe *both* in the correlation *and* in the existence of a mechanism that accounts for the observed correlation. Moreover, we hope we have laid to rest the worry about the psychological compellingness of mechanism-stories. We do not advocate accepting causal claims on the basis of stories about possible mechanisms: evidence of mechanisms is required, and good evidence at that.

§5
Conclusion

That correlation is insufficient for causation used to be a platitude. No longer. This is partly because our notion of correlation has become (slightly) more sophisticated: instead of simply considering whether the putative cause and the putative effect are probabilistically dependent, we now try to establish whether they are dependent conditional on other causes of the putative effect. The hope is then that evidence of this latter sort of correlation will be sufficient to establish causation. This hope has become so entrenched as to be built into explicit protocols for grading evidence.

In this paper we have argued that these explicit recommendations are wrong. For a wide variety of reasons, non-statistical evidence of mechanisms should often be used in conjunction with, rather than viewed as inferior to, statistical evidence of correlation. We need to revert to the dual methodology of Claude Bernard and Austin Bradford Hill, which considers evidence of mechansisms alongside evidence of correlation. This paper has aimed to provide some tentative first steps towards making the role of evidence of mechanisms more explicit.

We should emphasise that our argument is not with EBM as a general approach, nor with medicine as it is often practised. Our argument is with those protocols that undervalue non-statistical evidence of mechanisms. EBM has already seen a sequence of improvements to its evidence protocols—there is no reason why such protocols cannot be further refined to take proper account of the importance of evidence of mechanisms. Similarly, many in the medical community (including organisations, such as NICE and IARC, which have to make important public health recommendations) often give due weight to evidence of mechanisms in practice. The problem is that those who do so are frequently viewed as misguided on account of their going against the recommendations of explicit evidence hierarchies. It is because such hierarchies are prone to be followed uncritically that their improvement is such an urgent task.

Acknowledgements

We thank the UK Arts and Humanities Research Council for supporting this research. F. Russo also acknowledges financial support from the FWO-Flanders (2012-2013) as Pegasus Marie Curie Fellow. We are extremely grateful to the very many people who came to various events we organised during 2012, and participated in the discussions that allowed us to develop these ideas. We owe particular thanks to Ian McKay, Barbara Osimani, Jacob Stegenga and David Teira for extensive comments leading to significant improvements to the paper.

References

Bassler, D., Briel, M., Montori, V.M., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D. Walter, S.D., Guyatt, G.H., Flynn, D.N. et al. (2010). Stopping randomized trials early for benefit and estimation of treatment effects *Journal of the American Medical Association*, 303(12): 1180–1187.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical perspectives on cognitive neuroscience.* Routledge, Oxford.

Bell, G. (2008). *Selection: The Mechanism of Evolution, (2nd Edition).* OUP.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100:407–425.

Bernard, C. (1856). *An introduction to the study of experimental medicine.* Macmillan, New York, 1927 edition.

Bradford Hill, A. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300.

Bradford Hill, A. (1990). Memories of the British streptomycin trial. *Controlled Clinical Trials*, 2:77–79.

Broadbent, A. (2011). Inferring causation in epidemiology: mechanisms, black boxes, and contrasts. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the sciences*, pages 45–69. Oxford University Press, Oxford.

Brock, T. D. (1988). *Robert Koch. A Life in Medicine and Bacteriology.* ASM Press, 2 edition.

Campaner, R. (2011). Understanding mechanisms in the health sciences. *Theoretical Medicine and Bioethics*, 32:5–17.

Campaner, R. and Galavotti, M. C. (2012). Evidence and the assessment of causal relations in the health sciences. *International Studies in the Philosophy of Science*, 26(1):27–45.

Carter, K. C. (2012). *The Decline of Therapeutic Bloodletting and the Collapse of Traditional Medicine.* Transaction Publishers, New Brunswick.

Cartwright, N. (2007). Causal Powers: What Are They? Why Do We Need Them? What Can Be Done with Them and What Cannot? LSE discussion paper. http://www.lse.ac.uk/CPNSS/projects/CoreResearchProjects/ ContingencyDissentInScience/DP/CausalPowersMonographCartwrightPrint%20Numbers%20Corrected.pdf

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, 147:59–70.

Cartwright, N. and Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice*, 16:260–266.

Clarke, B. (2011a). *Causality in medicine with particular reference to the viral causation of cancers.* PhD thesis, Department of Science and Technology Studies, University College London, London.

Clarke, B. (2011b). Causation and melanoma classification. *Theoretical Medicine and Bioethics*, 32:19–32.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventative Medicine*, DOI 10.1016/j.ypmed.2012.10.020:in press.

Claveau, F. (2012). The Russo-Williamson theses in the social sciences: Causal inference drawing on two types of evidence. *Studies in History and Philosophy of Biological and Biomedical Sciences*, DOI 10.1016/j.shpsc.2012.05.004.

Cook, T. and Campbell, D. (1979). *Quasi-Experimentation. Design and Analysis Issues*

for Field Settings. Rand MacNally, Chicago.

Craver, C. (2007). *Explaining the Brain*. Clarendon Press, Oxford.

Cucherat, M., Haugh, M. C., Gooch, M., and Boissel, J.-P. (2000). Evidence of clinical efficacy of homeopathy: A meta-analysis of clinical trials. *European Journal of Clinical Pharmacology*, 56:27–33.

Daniels, M. and Bradford Hill, A. (1952). Chemotherapy of pulmonary tuberculosis in young adults. an analysis of the combined results of three medical research council trials. *British Medical Journal*, 31 May:1162–1168.

Darby, G. and Williamson, J. (2011). Imaging technology and the philosophy of causality. *Philosophy & Technology*, 24(2):115–136.

Darden, L. (2006). *Reasoning in Biological Discoveries*. Cambridge University Press, Cambridge.

Debré, P. (1994). *Louis Pasteur*. The John Hopkins University Press. English translation by Elborg Forster, 1998.

Demeulenaere, P., editor (2011). *Analytical Sociology and Social Mechanisms*. Cambridge University Press.

Dragulinescu, S. (2011). On 'stabilising' medical mechanisms, truth-makers and epistemic causality: a critique to Williamson and Russo's approach. *Synthese*, DOI 10.1007/s11229-011-0011-9.

Fayad, Z. A., Mani, V., Woodward, M., Kallend, D., Abt, M., Burgess, T., Fuster, V., Ballantyne, C. M., Stein, E. A., Tardif, J.-C., Rudd, J. H. F., Farkouh, M. E., and Tawakol, A. (2011). Safety and efficacy of dalcetrapib on atherosclerotic disease using novel non-invasive multimodality imaging (dal-PLAQUE): a randomised clinical trial. *The Lancet*, 378(9802):1547 – 1559.

Florey, M. E. (1961). *The Clinical Application of Antibiotics. Streptomycin and other Antibiotics Active against Tuberculosis*, volume 2. Oxford University Press, London, New York, Toronto.

Fox, W., Ellard, G., and Mitchison, D. (1999). Studies on the treatment of tuberculosis undertaken by the british medical research council tuberculosis units, 1946-1986, with relevant subsequent publications. *The International Journal of Tuberculosis and Lung Disease*, 3(10s2):S231–S279.

Gillies, D. A. (2011). The Russo-Williamson thesis and the question of whether smoking causes heart disease. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 110–125. Oxford University Press, Oxford.

Hedström, P. and Swedberg, R. (1988). *Social mechanism: An analytical approach to Social Theory*. Cambridge University Press.

Hitchcock, C. (2001). A tale of two effects. *The Philosophical Review*, 110(3):361–396.

Howick, J. (2011). Exposing the vanities—and a qualified defence—of mechanistic evidence in clinical decision-making. *Philosophy of Science*, 78(5):926–940. Proceedings of the Biennial PSA 2010.

IARC (2006). IARC monographs on the evaluation of carcinogenic risks to humans: Preamble. International Agency for Research on Cancer, http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf.

Illari, P. M. (2011). Mechanistic evidence: Disambiguating the Russo-Williamson thesis. *International Studies in the Philosophy of Science*, 25:139–157.

Illari, P. M. and Williamson, J. (2012). What is a mechanism? thinking about mechanisms *across* the sciences. *European Journal for Philosophy of Science*, 2:119–135.

Kshirsagar, A. V., Carpenter, M., Bang, H., Wyatt, S. B., and Colindres, R. E. (2006). Blood pressure usually considered normal is associated with an elevated risk of cardiovascular disease. *The American Journal of Medicine*, 119(2):133–141.

La Caze, A. (2008). Evidence-based medicine can't be ... *Social Epistemology*, 22(4):353–370.

La Caze, A. (2009). Evidence-based medicine must be ... *Journal of Medicine and Philosophy*, 34:509–527.

La Caze, A. (2011). The role of basic science in evidence-based medicine. *Biology and Philosophy*, 26(1):81–98.

La Caze, A., Djulbegovic, B., and Senn, S. (2012). What does randomisation achieve? *Evidence-based Medicine*, 17:1–2.

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.

Lucas, R. (1976). Econometric policy evaluation. In Brunner, K. and Meltzer, A., editors, *The Phillips curve and labor markets*, volume 1 of *Carnegie-Rochester Conference Series on Public Policy*, pages 161–168. North-Holland, Amsterdam.

Machamer, P., Darden, L., and Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, 67:1–25.

McLaren, L., Ghali, L. M., Lorenzetti, D., and Rockl, M. (2006). Out of context? Translating evidence from the North Karelia project over place and time. *Health Education Research*, 22(3):414–424.

MRC (1948). Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal*, 30 October:769–782.

MRC (1949). Treatment of pulmonary tuberculosis with para-aminosalicylic acid and streptomycin: Preliminary report. *British Medical Journal*, 31 December:1521.

MRC (1950). Treatment of pulmonary tuberculosis with streptomycin and para-amino-salicylic acid. *British Medical Journal*, 11 November:1073–1085.

Munro, S. A., Lewin, S. A., Smith, H. J., Engel, M. E., Fretheim, A., and Volmink, J. (2007). Patient adherence to tuberculosis treatment: A systematic review of qualitative research. *PLoS Med*, 4(7):e238.

NICE (2006). *The guidelines manual.* National Institute for Health and Clinical Excellence, London. Available from: www.nice.org.uk.

NICE (2009). *The guidelines manual.* National Institute for Health and Clinical Excellence, London. Available from: www.nice.org.uk.

NICE (2011a). *CG117: Tuberculosis.* National Institute for Health and Clinical Excellence, London. Available from: www.nice.org.uk.

NICE (2011b). *CG127: Hypertension: full guideline.* National Institute for Health and Clinical Excellence, London. Available from: www.nice.org.uk.

NICE (2011c). *CG127: Hypertension: quick reference guide.* National Institute for Health and Clinical Excellence, London. Available from: www.nice.org.uk.

Northcott, R. (2012). How necessary are randomized controlled trials? In Munson, R., editor, *Intervention and Reflection: Basic Issues in Medical Ethics*, pages 187–191. Thomson Wadsworth, 9th edition.

OCEBM Levels of Evidence Working Group (2011). The Oxford 2011 levels of evidence. Oxford Centre for Evidence-Based Medicine, http://www.cebm.net/index.aspx?o=5653.

Papineau, D. (1994). The virtues of randomization. *British Journal for the Philosophy of Science*, 45:437–450.

Puska, P., Vartiaine, E., Laatikainen, T., Jousilahti, P., and Paavola, M., editors (2009). *The North Karelia Project: from North Karelia to national action.* National Institute for Health and Welfare.

Reichenbach, H. (1949). *The Theory of Probability: An Inquiry into the Logical and*

Mathematical Foundations of the Calculus of Probability. University of California Press, Berkeley, CA.

Rothman, K., Greenland, S., and Lash, T. (2008). *Modern epidemiology*. Lippincott Williams & Wilkins.

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

Russo, F. and Williamson, J. (2011a). Epistemic causality and evidence-based medicine. *History and Philosophy of the Life Sciences*, 33(4):563–582.

Russo, F. and Williamson, J. (2011b). Generic versus single-case causality: the case of autopsy. *European Journal for Philosophy of Science*, 1(1):47–69.

Russo, F. and Williamson, J. (2012). EnviroGenomarkers: the interplay between mechanisms and difference making in establishing causal claims. *Medicine Studies*, 3(4):249–262.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72.

Salmon, W. C. (1977). Objectively homogeneous reference classes. *Synthese*, 36(4):399–414.

Salmon, W. C., Jeffrey, R. C., and Greeno, J. G. (1971). *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, Pittsburgh, PA.

Schwartz, G. G., Olsson, A. G., Abt, M., Ballantyne, C. M., Barter, P. J., Brumm, J., Chaitman, B. R., Holme, I. M., Kallend, D., Leiter, L. A., Leitersdorf, E., McMurray, J. J., Mundl, H., Nicholls, S. J., Shah, P. K., Tardif, J.-C., and Wright, R. S. (2012). Effects of dalcetrapib in patients with a recent acute coronary syndrome. *New England Journal of Medicine*, 367(22):2089–2099. PMID: 23126252.

Sober, E. (1988). The principle of the common cause. In Fetzer, J. H., editor, *Probability and causality: essays in honour of Wesley C. Salmon*, pages 211–228. Reidel, Dordrecht.

Solomon, M. (2011). Just a paradigm: evidence-based medicine in epistemological context. *European Journal for Philosophy of Science*, 1:451–466.

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

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences*, 42:497–507.

Stein, E. A., Roth, E. M., Rhyne, J. M., Burgess, T., Kallend, D., and Robinson, J. G. (2010). Safety and tolerability of dalcetrapib (RO4607381/JTT-705): results from a 48-week trial. *European Heart Journal*, 31(4):480–488.

Thompson, R. P. (2011). Causality, theories and medicine. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 25–44. Oxford University Press, Oxford.

Timmermans, S. and Berg, M. (2003). *The gold standard: The challenge of evidence-based medicine and standardization in health care* Temple University Press, Philadelphia.

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

Wagner, E. H. (1982). The North Karelia Project: what it tells us about the prevention of cardiovascular disease. *American Journal of Public Health*, 72(1):51–53.

Weber, E. (2009). How probabilistic causation can account for the use of mechanistic evidence. *International Studies in the Philosophy of Science*, 23(3):277–295.

Williamson, J. (2005). *Bayesian nets and causality: philosophical and computational*

foundations. Oxford University Press, Oxford.
Williamson, J. (2013). How can causal explanations explain? *Erkenntnis*, in press.
Worrall, J. (2002). *What* evidence in evidence-based medicine? *Philosophy of Science*, 69:S316–S330.
Worrall, J. (2007). Why there's no cause to randomize. *British Journal for the Philosophy of Science*, 58:451–488.
Worrall, J. (2010). Evidence: philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice*, 16:356–362.