Evidence evaluation and the epistemology of causality in medicine

Daniel Auker-Howlett

Department of Philosophy

University of Kent

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

2020

Word count: 74,592

I, Daniel Auker-Howlett, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Date: 06/04/2020

Signature:

Contents

Acknowledgements

Firstly, I would like to thank my co-supervisors, Jon Williamson and Kristoffer Ahlstrom-Vij. Their commitment to seeing my work improve, as well as their support outside of my primary research, has been instrumental to my academic development. The vast majority of the ideas in this thesis were developed during and in response to their patient questioning of my work over the past three years. I would also like to thank Graeme A. Forbes for stepping in as a third supervisor during the latter part of the process, and for helping me clarify the structure of the thesis.

Special thanks must go to Michael Wilde for support both academic and personal throughout the PhD. I also thank Michael, and William Levack-Payne, for proof-reading portions of this thesis, and for providing valuable advice that helped refine the end product.

I would also like to recognise the environment of the philosophy department at Kent for the support it has offered during this time. In particular, I would like to thank all those who have attended weekly 'evidence seminars', especially the medical nihilism reading group, where I have developed and honed ideas. For personal support, I have to show my gratitude to all members of the Kent philosophy post-graduate community, as they have been invaluable for keeping my spirits up during what is not always an easy time.

Finally, my thanks to my friends and family, without whom I would not have got to this point. To my friends, Tim and Joe, and my brothers, Connor and Kieran, for continual support during the PhD, and keeping me in beer during the writing up period. And most of all, to my Mum and Dad, for never ceasing to support me in all my endeavours.

Preface

Establishing causal claims is important across medicine. For example, a new drug is effective at lowering blood pressure if and only if the drug is a cause of a decrease in blood pressure. To work out if the drug does indeed cause a decrease in blood pressure requires a methodology for causal evaluation. The dominant methodology for causal evaluation in medicine is that of Evidence Based Medicine (EBM). A competing methodology is EBM+. The aim of this thesis is to defend EBM+ as a more complete epistemology of causality in medicine than that of EBM.

An epistemology of causality in medicine must answer two questions: i) what kinds of evidence should be used to evaluate causal claims?; ii) how should that evidence be evaluated? On the one hand, EBM and EBM+ differ on i), as EBM holds that only evidence from clinical studies can be used to evaluate causation, whereas EBM+ admits evidence from both clinical and mechanistic studies, which utilise the methods of the biomedical sciences. On the other hand, they mostly agree on ii), as they agree that evidence should be evaluated explicitly and transparently. Where they do disagree on ii), they do so only because they disagree on what kinds of evidence should be evaluated. What counts as evidence for causality is thus the key difference between the two methodologies.

The plan of the thesis is as follows. In Chapter 1 I introduce both EBM and

EBM+. I then set out the conceptual framework I use, as well as introduce the case studies I analyse, throughout the thesis. I defend an *evidentially pluralist* position on what counts as evidence for causation in Chapter 2. EBM+ is evidentially pluralist and so is provisionally a closer approximation to a complete epistemology of causality in medicine than EBM. It is not enough to just be evidentially pluralist when it comes to evaluating causality, as EBM+ must also be able to provide methods for the evaluation of evidence of mechanism, particularly that obtained from mechanistic studies. In the remainder of the thesis I analyse an evaluative framework, built on EBM+ principles, namely, ‘Evaluating evidence of mechanisms in medicine’ (EEMM) (Parkkinen et al., 2018b).

In general, one can have feasibility, practical, conceptual and malleability worries about any methodology of causal evaluation. To allay feasibility worries, I use EEMM to carry out a systematic review of evidence of mechanism obtained from mechanistic studies in Chapter 3. I then defend EEMM against some practical worries about establishing mechanisms on the basis of mechanistic studies in Chapter 4. A conceptual worry about the evaluative process in EEMM concerns how to characterise judgements about the strength of evidence. In particular, in Chapter 5 I defend the interpretation used in EEMM of ‘quality of evidence’, which is put in terms of ‘stability of confidence’. Malleability worries concern the role of expert judgement in evidence evaluation: whether subjective choices influence judgements to an inappropriate extent. This worry is motivated by the *medical nihilism* thesis, which claims that we should only ever hold low confidence in the effectiveness of medical interventions. As medical nihilism poses a problem to any methodology for causal evaluation in medicine, I reject it in Chapter 6. I do however concede that the need to make subjective choices during evaluations makes the malleability of methods a worry that any evaluative framework must contend with. Accordingly, in

Chapter 7 I consider the potential to use formal models of belief to constrain the influence of subjectivity on judgements. I then use lessons learnt in Chapters 5 and 7 to ameliorate a potential example of malleability in EEMM, by formulating additional guidance in Chapter 8. I motivate further extensions to the EEMM framework in Chapter 9. Such extensions involve integrating it with a clinical study evaluator, the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) framework. I finish by summarising the claims of this thesis, drawing consequences for real world practice, and identifying open questions.

I conclude that EBM+ as a methodology of causal evaluation is feasible, conceptually defensible, and can counter malleability concerns. Moreover, where appropriate I propose recommendations on how to improve on the EBM+ analysis and the EEMM methodology.

Parts of this thesis are derived from, or were developed during the production of, previously published articles. A version of the introduction to EEMM found in §3.2 can be found in Abdin et al. (2019, Section 4), to which I was the sole contributor of that section. In Chapter 2, I appeal to, rather than alter and include, ideas found in papers to which I was a contributor: Where relevant, I cite Auker-Howlett and Wilde (2019) and Auker-Howlett and Wilde (2020), as the research and analysis of the case study found in this article was carried out in collaboration with Michael Wilde.

List of Figures

List of Tables

Part I

Evaluating causal claims in medicine

Chapter 1

Introduction

1.1 Epistemologies of causality in medicine

In this section I introduce two competing epistemologies of causality in medicine. The point at which they differ is on what counts as evidence for causation.

1.1.1 EBM

A long tradition in the philosophy of science has been concerned with what counts as good evidence (e.g., Mayo (1996); Achinstein (2001)). Such concerns are important in medicine as well, as it is plausible that there will be some kinds of evidence that can establish causal claims in medicine, and some that cannot. Moreover, between those that can, some will be better at it than others. EBM holds that the best evidence for causal evaluation is obtained from *clinical studies*, where clinical studies employ methods that test the claim that A causes B by "repeatedly measur[ing] values of a set of measured variables that includes the variables A and B" (Parkkinen et al., 2018b, p.14). Accordingly, clinical studies observe associations between A and B. This can establish that A and B are *correlated*, where to be correlated is for B to be probabilistically dependent on A, such that $P(B|A) > P(B)$. This means that the probability of B is raised when A also obtains. Inferences to the claim that A *causes* B are then warranted by the kind of method employed in the study. Warrant for inferring that a correlation is causal in nature is provided by ruling out plausible alternative hypotheses for the correlation (Howick, 2011b, pp.33-38). The extent to which different methods can achieve this aim is variable, which introduces a hierarchy of clinical study methods.

The "gold standard" methodology within EBM is the Randomised Controlled Trial (RCT). RCTs are *experimental* studies, as the researcher places participants into intervention or control groups. They then manipulate exposure to the putative cause by selectively administering it to only the intervention group. Differences in the clinical outcome between the two groups are then measured to work out if the intervention is correlated with the outcome. A part of the warrant for inferring causation from a correlation is *randomisation*, where participants in a trial are randomly assigned to the intervention and control groups. The purpose of this technique is to equally balance out all potential *confounders*, where a confounder is a variable that is correlated with both the putative cause and effect, and is the true cause of the effect. A potential confounder is any variable that is correlated with both the putative cause and effect and could account for some variation in the effect. Some potential confounders will be made *plausible* by evidence relevant to the area under investigation. For example, for many mild infectious diseases, age is often a cause of recovery. To be precise, young adults have more robust immune systems than older adults do, and so are more likely to recover without intervention. If a study administers a novel anti-viral to only young people with a mild viral infection, and

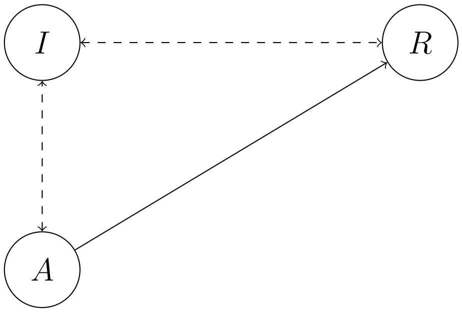


Figure 1.1: Causal diagram for confounding by age (A) of the correlation between intervention (I) and recovery (R). Dotted lines indicate correlation; bold line indicates causal relation.

observes a correlation between the intervention and recovery, then this correlation may be attributable to the age of the participants rather than the anti-viral. This is because age is correlated with both administration of the intervention and with recovery, and is known to cause recovery (see fig. 1.1 for a visualisation of these relationships). To balance out this plausible confounder, a researcher could intentionally assign participants to trial groups, such that the ages of participants in each group are balanced (e.g., both groups have 10 participants in the 18-26 range, 10 participants in the 18-34 range, and so on). However, there are indefinitely many potential confounders that are not plausible given our current evidence. They are called *unknown* confounders (see fig. 1.2). The problem is that researchers cannot intentionally balance unknown confounders across groups. Randomly assigning participants to control groups is arguably a way to balance both plausible and unknown confounders. The effect of any confounders in the intervention group should then cancel out their effects in the control group. A correlation observed between the intervention and clinical outcome can thus be attributed to a causal relationship holding, rather than confounding.

With randomisation alone, one cannot infer causation from an observed correlation. While the effect may not be the result of a confounder, it is still possible it is the result of the *placebo effect*, a phenomenon that involves the elicitation of par-

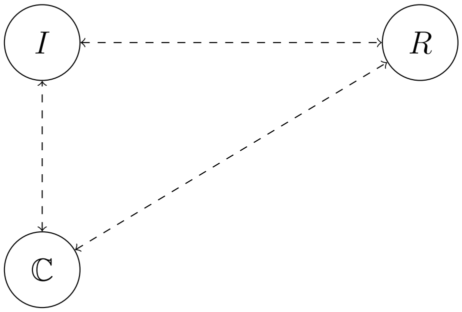


Figure 1.2: Causal diagram for confounding by the set of unknown confounders $\mathbb{C}$ of the correlation between intervention (I) and recovery (R). $\mathbb{C}$ is a potentially infinite set of unknown confounders. Dotted lines indicate correlation. It is unclear whether any $C_i \in \mathbb{C}$ *is* a cause of R, so the relationship is defined as correlation. One cannot infer causation between I and R because the relationship between any C_i and R could be causal.

ticipants' own internal self-healing processes solely by the process of administering treatment. The experimental nature of the study design allows the use of *placebo controls*, where a placebo control mimics the nature of the administration, but not the supposed clinically effective components of the intervention. For example, a sugar pill in place of a pill containing propanolol, which is a β-blocker used to lower blood pressure. Employing the placebo control allows the subtraction of the extent of an observed effect that is attributable to the placebo effect, from the difference between control groups (see fig. 1.3). If there is still a correlation between putative cause and effect, then one has ruled out the alternative 'placebo hypothesis'. To ensure that the placebo control works correctly, a technique called *double-blinding* is implemented. A study is double-blinded when neither the participants nor the researchers know which study group is in the intervention or control group. This prevents researchers from giving preferential treatment to a study group as a result of knowing whether they will receive the potentially effective intervention. It also prevents differences in expectations of participants about whether they will benefit or not from the 'treatment'. Either of these factors can elicit the placebo effect. Double-blinding is also used to rule out a range of other explanations, e.g., a patient that realises they are in the control group may start self-medicating. For a full description and justification of these techniques, see Howick (2011b, pp.33-116).

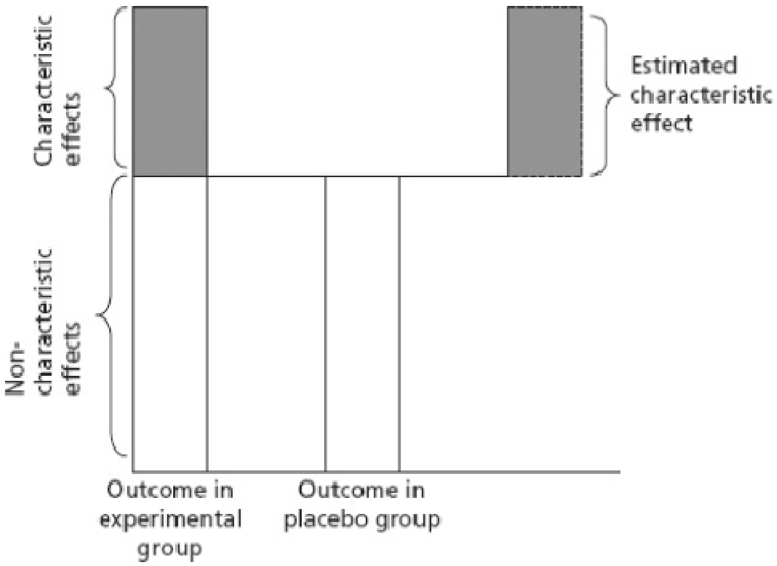

Characteristic effects
Non-characteristic effects
Estimated characteristic effect
Outcome in experimental group
Outcome in placebo group

Figure 1.3: Graph demonstrating how placebo effect subtracted from total effect gives characteristic effect, from Howick (2011b). The characteristic effect is the effect of the supposed clinically effective component of the intervention.

Evidence obtained from an RCT is typically considered better for causal evaluation than from studies whose design is non-randomised, which can be experimental or *observational*. Experimental non-randomised studies measure comparisons of outcomes between intervention and control groups, and can implement placebo controls and double-blinding. Nevertheless, they may struggle to rule out confounding on account of study design alone. Observational studies only measure exposure to the putative cause rather than selectively administering it to only the intervention groups. Examples of such studies include *case control studies*, and *cohort studies*. Case-control studies identify groups of patients who display the putative effect, match them with groups of patients who do not display the effect, and then researchers identify whether those patients were exposed to the intervention or not. Cohort studies take the opposite approach and identify groups that have or have not been exposed to a putative cause, and identify whether there is a difference in outcomes between the groups. In both observational study-designs, associations can be identified between exposure and outcome, but it is difficult to infer causation on account of study-design alone. This is again because of the lack of a technique for ruling out potential confounders. So, randomised studies are typically considered better methods for evaluating causality. This characterisation is rather coarse, as there are techniques available to researchers to rule out alternative explanations when making inferences from the results of non-randomised studies. This point will be picked up on in more detail in later chapters, but it suffices here to say that these techniques are not a standard part of the *design* of those methods, even if they are used when making causal inferences. Hence, the hierarchy of study designs within EBM.

Before moving on to critiques and proposed developments of EBM, some clarifications are in order. Firstly, when the domain of investigation is not clinical science, reference is made to *epidemiological studies* providing the best evidence. This will be the case in exposure assessment, e.g., the assessment of what compounds are carcinogenic by the International Agency for Research on Cancer (IARC). The designs in both domains will however be broadly the same, if not often identical. In the remainder of this thesis I will refer to clinical studies, but this usage is meant to include any study design that "repeatedly measures values of a set of measured variables that includes the variables A and B" (Parkkinen et al., 2018b, p.14), and employs particular techniques, the purpose of which is to warrant inferences from observed correlations to causation. Note also that RCTs do not strictly provide the best evidence for causal evaluation. Systematic reviews and meta-analyses of RCTs are typically considered to provide better evidence than RCTs alone. Such methods aggregate and evaluate the evidence from RCTs to arguably give better estimates of effects than single studies can alone. But they are statistical *evidence aggregation* methods, not clinical studies. And what they aggregate is evidence from clinical studies. I therefore maintain that the EBM position is that randomised study designs provide better evidence than non-randomised designs, and all comparative clinical studies provide better evidence than any other kind of study methodology.

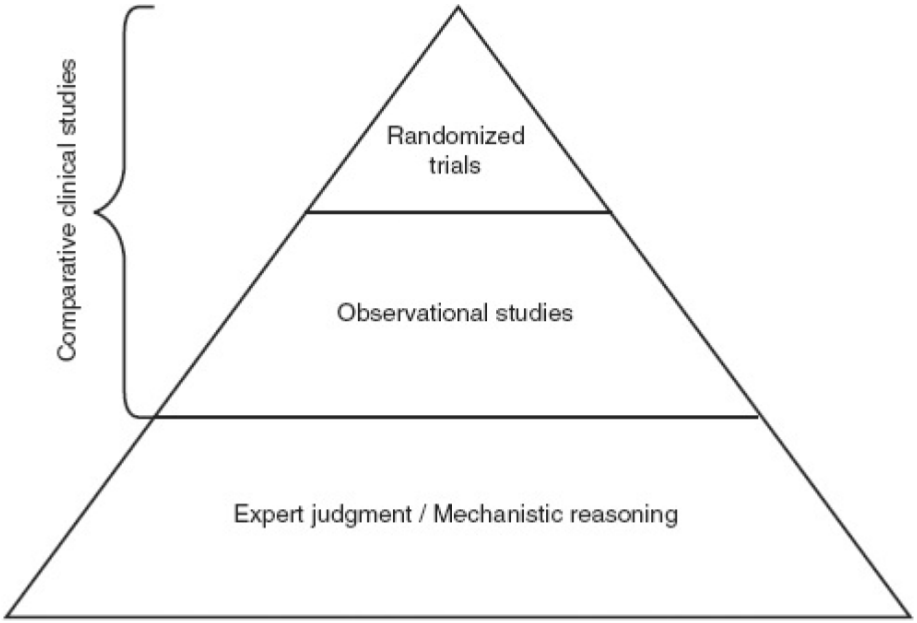

Comparative clinical studies
Randomized trials
Observational studies
Expert judgment / Mechanistic reasoning

Figure 1.4: Hierarchy of study-designs from (Howick, 2011b, p.5). He notes in the caption for this figure that "systematic reviews of all study-designs [are] assumed to be superior to single studies".

This kind of view is supported by the hierarchy of study-designs found in Howick (2011b) (see fig. 1.4), who notes that, for all designs, aggregation of multiple studies will provide better evidence than one study alone.

1.1.2 EBM+

Many criticisms of EBM focus on the claim that some kinds of evidence are better than others for evaluating causality. For example, there have been critiques of the claim that RCTs provide the best evidence (Worrall, 2002, 2007; Grossman and Mackenzie, 2005), as well as critiques of the ability of RCTs to establish how well an intervention works outside of the trial population (Feinstein and Horwitz, 1997; Cartwright, 2007). As one example, John Worrall (2002, 2007) objects to the view that randomisation controls for all confounders known and unknown, by balancing those confounders across the treatment and control group. He argues that there is still a chance that a randomised study has unbalanced groups. Given that there are potentially an infinite number of unknown confounders, in any particular random allocation of participants the chance that there is an imbalance in confounders between groups is high. Proponents may accept this, but still maintain that randomisation *tends to balance* confounders. The problem with this, so Worrall's argument goes, is that 'tends to balance' refers to repeating the trial indefinitely with a new random allocation in each repetition. As the number of repetitions increases, the probability that confounders are balanced also increases. However, it is the case that any given RCT has not been performed indefinitely. We therefore cannot be sure that the groups are balanced for all confounders, known and unknown. His proposal to deal with this issue is to reject the hierarchy of study designs. In particular, he argues that observational studies that match known confounders are no worse for inferring causation than randomised studies.

Another proposal to deal with the problems facing standard EBM is to take seriously the role that mechanisms can play in causal evaluation. Recent developments in medical methodology have highlighted the role that mechanisms may play in improving on the methods of causal evaluation in medicine (Clarke et al., 2013, 2014; Parkkinen et al., 2018b). *Mechanistic theories of causality* analyse causality as: A causes B if and only if there is an appropriate mechanism linking A and B (Gillies, 2019a, p.72). There are many kinds of mechanistic theories of causality (see (Gillies, 2019a, pp.72-79) for an overview). Each theory says something about the nature of causality and there are a number of different characterisations of what a mechanism is (e.g., Machamer et al. (2000); Glennan (2002); Bechtel and Abrahamsen (2005); Illari and Williamson (2012)). However, the motivation for including mechanisms in causal evaluation is epistemological in nature, rather than metaphysical.

The EBM+ proposal is to include explicit evaluation of evidence of mechanisms alongside and on a par with evidence of correlation. The motivation for this development is an epistemological thesis made by Russo and Williamson (2007): in order to establish a causal claim (A causes B) in medicine, one normally needs to establish that A is appropriately correlated with B, and establish that a mechanism exists linking A and B that can account for the correlation. This thesis was dubbed the Russo-Williamson Thesis (RWT) by Donald Gillies (2011) and I follow this usage. What it takes for A and B to be correlated is as stipulated above. There is a mechanism linking A and B when there exists a sequence of features between A and B that explains instances of B in terms of instances of A (Parkkinen et al., 2018b). Those features can be entities (e.g., proteins), activities (e.g., binding), or the way the mechanism is organised. As the organisation can be linear or highly complex, so can the sequence of features linking A and B. The idea is that inferring causation from evidence of correlation or mechanism alone faces distinct problems, but each kind of evidence can help to rule out the problems facing the other kind. Evidence of mechanisms helps to rule out confounding, which is the potential error facing inferring causation from an observed correlation. Mechanisms alone cannot establish causation because even if it can be established that there is a mechanism linking A and B, other mechanisms may also be operating that prevent B from occurring. This is the problem of masking, which precludes inferring from the fact that a mechanism operates between A and B, to the conclusion that B will be brought about by A. Finding that a correlation exists between A and B establishes that when A is present the probability of the effect B is raised. This is often put in terms of showing that there is a *net effect* of A, as any counteracting mechanisms do not suffice to completely cancel out B. So to establish causation we need to rule out errors stemming from both confounding and masking, which requires establishing both a correlation and a mechanism.

EBM and EBM+ are competing epistemologies of causality in medicine because they answer differently the question of what kinds of evidence should be used to evaluate causal claims. Evidence that can establish a mechanism is typically obtained from *mechanistic studies*, where mechanistic studies provide evidence for the claim that A causes B by providing evidence of the features of a mechanism linking A and B (Parkkinen et al., 2018b, p.14). The methods employed by mechanistic studies are various, but they are the methods of the "basic sciences" (e.g., microbiology, biochemistry, physiology, molecular biology etc.). In some versions of EBM, evidence from mechanistic studies is not considered as evidence for causality. Where it is considered evidence, it is always taken as strictly inferior to evidence obtained from clinical studies. Reasons for this will be explored in Chapter 2, but it suffices to say here that EBM and EBM+ clash at least on what counts as evidence for causality. However, characterising the clash as one of evidence from clinical studies versus evidence from mechanistic studies is only a first approximation.

1.2 Distinctions and disambiguation

To make sense of the key differences between EBM and EBM+ requires being clear on the way in which we distinguish between kinds of evidence. Contrasting EBM and EBM+ in terms of the kinds of studies that can provide evidence of causation may not be warranted. In this section I introduce a better way of distinguishing between kinds of evidence. This will allow a more precise delineation of the competing views on what counts as evidence for causation.

First note that there is a discordance between the characterisation of the two epistemologies of causation in medicine. EBM is typically characterised in terms of preferring one method for obtaining evidence over another. On the other hand, while EBM+ seeks to add to EBM the evaluation of evidence obtained from mechanistic studies, the thesis that motivates it, RWT, is characterised in terms of the objects of evidence, namely, correlations and mechanisms. Phrasing in terms of objects of evidence is derived from a distinction made by Phyllis Illari (2011) between the methods by which evidence is obtained, and the objects of that evidence. For example, an RCT is a type of *evidence-gathering method*, and evidence can be *of* a variety of things: of cancer; of heart disease; of DNA. Gillies (2019a) calls this *Illari's distinction* and I follow this usage. This distinction is used to *disambiguate* RWT. The ambiguity in the original formulation of RWT was that one can interpret its terms as saying there are either: i) two kinds of evidence-gathering methods typically necessary for establishing causation, namely *mechanistic evidence-gathering methods* and *correlational evidence-gathering methods*; ii) two kinds of objects of which one needs evidence, if one is to establish causation, namely, evidence of a mechanism and evidence of a correlation. Illari (2011) rules out the first interpretation, and argues the second one actually tracks a useful distinction.

I will detail Illari's arguments below because the conceptual framework of this thesis makes use of Illari's distinction in a number of ways. Firstly, to properly contrast EBM and EBM+, Illari's distinction needs to be applied to EBM as well, which I argue for in this section. Secondly, debates over the kinds of evidence admitted to causal evaluations can be clarified when this common conceptual ground is set out (Chapter 2). The evidence review conducted in Chapter 3 makes use of the distinction by evaluating the claim *that a mechanism exists* using evidence from both mechanistic and clinical studies. Finally, some methodological problems facing evidence evaluation can be better addressed within this framework (Chapter 4). As Illari's aim was to *disambiguate* RWT, I will refer in the remainder of this thesis to arguments that rely on this distinction as the *disambiguation approach*. First, some clarifications are in order.

The original formulation of RWT was put in terms of *mechanistic evidence* and *probabilistic evidence* (Russo and Williamson, 2007). Instead of probabilistic evidence, Illari chooses to use the term *difference-making evidence*, where a causal conclusion is reached by obtaining evidence that a cause makes a difference to an effect. Illari chooses this term because it is both a broader category that includes things like correlations and counterfactual relationships (if I had not thrown the ball, the window would not have broken), and because Russo and Williamson claim that probabilistic evidence is used to show that the cause makes a difference to the effect. To complicate matters, the most recent formulation of RWT refers to correlation instead of either probabilities or difference-making (Parkkinen et al., 2018b; Williamson, 2019). In my exposition of the arguments for Illari's distinction I will refer to correlations where she refers to difference-making. This is for two reasons. One is that, in medicine, the difference-making relations we are concerned with are typically correlations. The other is that the most recent formulations of RWT are included in the evaluative framework built on EBM+ principles that I analyse in the bulk of this thesis (introduced in Chapter 3). Referring to correlations will enable consistency of usage throughout this thesis. Now to Illari's arguments.

Firstly, types of evidence are in general distinguished by properties that are relevant to conclusions. One simple way this works for evidence-gathering methods is distinguishing by *tools and techniques*: each kind of technique will have strengths and weaknesses, and those limitations allow effective understanding of the techniques and the conclusions they can support. Illari provides the example of different kinds of spectroscopy, that give different kinds of information:

> "Infrared absorption spectroscopy and Raman spectroscopy both give information about vibrational modes of molecules, but are complementary because the modes of vibration that show up by infrared don't show up

> by Raman, and vice versa." (2011, p.142)

Because of the differences in the kinds of conclusions possible on each kind of method, one can distinguish between types of evidence that come from infrared spectroscopy and from Raman spectroscopy. One can also distinguish at higher levels of abstraction. Illari offers three ways of distinguishing methods in this manner: Quantitative v. qualitative; generic v. single case; evidence that needs repeated trials v. evidence that can confirm on one or a few experiments. Again, each kind of evidence tracks properties relevant to a different kind of conclusion. One can also combine the distinctions in order to further distinguish kinds of evidence. For example, a clinical trial might provide quantitative and generic evidence obtained by means of repeated trials. This will support a quantitative, generic causal conclusion: in population N, A causes B in x% of p. If the evidence were single case, it would support a conclusion about a specific person from N. Evidence that could confirm a hypothesis on only one or a few experiments is likely not to be obtained using a clinical trial. Such evidence must enable one to "see pretty straightforwardly that something must be the case" (Illari, 2011, p.143), and is typically only provided by imaging experiments that allow one to see biochemical structure.

Illari argues that because of the link between evidence and conclusions, 'tools and techniques' and the three more abstract categories are all useful ways of distinguishing between types of evidence-generating methods. One might be tempted to categorise mechanistic evidence-gathering methods and correlational evidence-gathering methods as a specific combination of the less abstract, but useful, categories of method. For example, mechanistic evidence-gathering methods could be seen as qualitative, single case and requiring one or few experiments to confirm. This might be the case if imaging technology is used to provide evidence of the structure of a protein that plays a crucial role in the operation of a mechanism. One can pretty straightforwardly see the structure of the protein, which means one has qualitative evidence of this protein and one does not need to repeat the experiment many times to get this evidence (although Illari notes that one may need to repeat the experiment to rule out bias and fraud). However, distinguishing evidence-gathering methods by combinations of these properties is not possible. While some evidence of mechanism may be obtained in methods that provide quantitative, single-case evidence, it is also the case that evidence of mechanism can be obtained by repeated experimental interventions, as was the case with Crick and Brenner's work on cracking the genetic code (Illari, 2011, p.143). This type of method is also a standard way of obtaining evidence of correlation. Moreover, even the same kinds of experimental methods can be used to obtain both evidence of mechanism and evidence of correlation. For example, there is no reason why a clinical trial cannot provide evidence of a mechanism. This all means that the link is severed between conclusions and methods: there seems to be no way of dividing up methods into those that provide evidence of mechanism and those that provide evidence of correlation. Therefore, there is no principled distinction one can make between mechanistic evidence-generating methods and correlational evidence-generating methods.

On the other hand, there *is* a principled reason for distinguishing types of evidence by the kind of object it is evidence of. The reason is that, similarly to the useful types of evidence-generating methods, the objects of evidence have particular strengths and weaknesses when it comes to inferring causation on the basis of that evidence. And those strengths and weaknesses complement one another. This is the basic motivation for RWT and is outlined above. So the objects of evidence have properties that are relevant to the kinds of conclusions they support, namely the complementary strengths and weaknesses of each kind of object. This is important, as if one did not distinguish between objects of evidence then one would miss out on those properties that are relevant to causal conclusions. Thus, one can and should distinguish between objects of evidence. This leads to a requirement that, to be defensible, RWT should be formulated in terms of evidence of mechanism and evidence of correlation.

The distinction can be applied to the EBM view on evidence, which holds that the best evidence is that obtained by comparative clinical studies. Due to Illari's distinction, one cannot characterise this view as that correlational evidence-gathering methods provide the best evidence. Because clinical studies can provide evidence of both correlation and mechanism, one might not be able to say that EBM sees evidence of correlation as the best kind of evidence. A proponent of EBM might say instead that the object of evidence is causation. This is because the EBM stance on evidence is justified by appeal to how well a method can exclude bias, particularly bias due to confounding (Howick, 2011b). Methods that are designed to rule out more confounding will be better, *ceteris paribus*, than those that rule out less. And this is because ruling out confounding allows one to infer causation. So, on the EBM view, the reason why one method is ranked higher than another is its ability to obtain better evidence of causation.

However, this would not be direct evidence of causation. Instead, one infers causation by ruling out confounding. And when one is making an inference to causation by ruling out confounding, the object of evidence is a correlation. So the EBM view should be characterised as that evidence of correlation from clinical studies is the best kind of evidence for evaluating causality. One might think to the contrary causation is directly inferred from results obtained by the best kinds of clinical study. It is true that inferring causation from evidence obtained in an RCT may not always involve an explicit process of inferring causation on the basis of observing a correlation, if its techniques for ruling out confounding are implemented correctly. However, it is possible on some EBM views to infer causation from evidence obtained in observational studies. And in those cases, a more explicit process of inferring from a correlation to causation is carried out, e.g., the systematic ruling out of plausible confounders through statistical analysis. Whether it be implicit or explicit, there is still a process of inferring from correlation to causation. Indeed, Miriam Solomon, a proponent of the EBM view, says that

> “[h]ealth care interventions are judged effective when there is a correlation between the intervention and positive outcomes. Often it is not too much of a leap to infer that the intervention causes the positive outcome.” (2015, p.117)

A final problem for my characterisation of the EBM view is that it is also possible that a clinical study could provide evidence of mechanism. But one would not infer causation from evidence of mechanism through a process of ruling out confounding. So the justification for inferring causation from evidence obtained in clinical studies is not relevant to inferring causation from evidence of mechanism. Therefore, the EBM view on evidence is that evidence of correlation from clinical studies is the best kind of evidence for evaluating causality.

There are however nuances between different versions of this view. I will go in more detail into the different kinds of rejections of evidence of mechanism in Chapter 2, but EBM views can be distinguished thus. Some EBM proponents hold that only evidence of correlation from clinical studies can be evidence of causation. For example, the GRADE framework for evaluating causality (Balshem et al., 2011) does not consider evidence of mechanism at all. This sort of view is *evidentially monist*. Solomon (2015) thinks that at best, even good evidence of mechanism is only weak evidence of causation. So causal claims can only be established on the basis of (good) evidence of correlation. Others hold that evidence of correlation can overturn evidence of mechanism in the case of conflict (e.g. Howick (2011b)). What ties these kinds of views together is that good evidence of correlation from clinical studies is superior to good evidence of mechanism. Hence, I call holders of such views, proponents of *evidential superiority*.

On the other hand, EBM+ considers no kind of evidence to be superior to another, as long as when evaluating causality one can establish both the existence of a correlation and of a mechanism. By not privileging one kind of evidence over another, EBM+ is *evidentially diverse*. There are however a number of different kinds of evidential diversity. Firstly, proponents of RWT cast their stance as *evidential pluralism*, as in this quote from (Parkkinen et al., 2018b, p.4):

> "In the philosophy of causality, the following thesis has been put forward (Russo and Williamson, 2007): **Evidential Pluralism** - This is the thesis that one typically needs both evidence of correlation and evidence of mechanisms to establish a causal claim."

Evidential pluralism is not always equated with RWT. As Stefano Canali notes: "Federica Russo and Jon Williamson introduced *a version* of evidential pluralism" (2019, p.4, *my emphasis*), that version being RWT. To add clarity to the debates, I will call the position based on RWT 'evidential pluralism'. Positions that are evidentially diverse are those where no one kind of evidence is superior to another. One key difference between evidential pluralism and other kinds of evidential diversity is whether any one kind of evidence is *necessary*. For example, Gillies (2019a) defends a weaker version of RWT that does not make establishing a mechanism necessary for establishing causality. I address such positions in §2.3. Note that there are many commonalities between broad evidential diversity and evidential superiority. Both admit many kinds of evidence. The difference between them is that evidentially diverse views would not say that good evidence of correlation is better than good evidence of mechanism. That a view could fall easily into both the evidential superiority and one of the broader evidential diversity categories is no matter for the claims made in the next chapter, as they are united in their rejection of evidential pluralism.

In sum, EBM+ is both evidentially diverse and pluralist, requiring evidence of correlation and evidence of mechanism in order to establish causation. Common to all kinds of EBM is that evidence of mechanism is not required in order to establish causation. Hence, while EBM may accommodate an evidentially diverse epistemology, it cannot accommodate a pluralist one. On this point of difference, a supporter of RWT would conclude that "[c]urrent EBM provides a reasonable first approximation to the correct epistemology", but "EBM+ provides a better approximation" (Williamson, 2019, p.42). In other words, EBM+ is a more complete epistemology of medicine, the thesis I intend to defend.

1.3 Case studies

In this thesis I will use three main case studies. The evaluative framework built on EBM+ principles, EEMM, and a case study that concerns the severe respiratory disease 'Middle East respiratory syndrome' (MERS), are introduced in full in Chapter 3. I use them both to analyse evidence evaluation on EBM+. Throughout this thesis I will also appeal to an example of the EBM approach to causal and evidence evaluation. This is the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) framework. I utilise GRADE in a number of ways. Broadly, I use it to compare the EBM+ approach to evidence evaluation to, as well as to provide evidence for claims about evidence evaluation in general. At points I will also analyse it, defending it against criticisms, and making suggestions for improvement, where appropriate. As I refer to GRADE at points in the thesis that occur before I carry out any kind of analysis of it, to aid with the flow of the thesis I introduce the framework in full here.

GRADE is a framework for assessing the quality of evidence in clinical research. To this end, GRADE provides a step-by-step procedure for decision makers to follow, such that the process can output a clear rating of the 'quality of evidence'. The transparency and ease of use of this procedure is its strength, and has seen its uptake by important public health organisations such as the World Health Organisation (WHO), the National Institute for Clinical Excellence (NICE) in the UK, the Cochrane Collaboration, and the Agency for Healthcare Research and Quality (AHRQ) in the USA. When I refer to the *framework* I refer to the entirety of the GRADE approach, as set out in the numerous publications by the GRADE *working group* (e.g., Guyatt et al. (2011a); Balshem et al. (2011); Hultcrantz et al. (2017)). The working group is composed of experts in the field of clinical science from throughout the world, and is responsible for designing, updating, and communicating the GRADE framework.

The framework itself provides a step-by-step procedure that outputs ratings of quality of evidence, at levels: High; Moderate; Low; Very-Low. The interpretation of a quality rating is that it "reflects the extent of our confidence that the estimate of effect is correct" (Balshem et al., 2011, p.403). The procedure is composed of two stages. In Stage 1, users are instructed to consider the methodology of the study under consideration: If the study is a randomised design then it is given an initial high-quality rating; if it is a non-randomised design then it is given an initial low-quality rating. In Stage 2, users are instructed to consider a number of *moderating domains*, where moderating domains are broad categories of features of evidence that can alter quality ratings. While there are domains for up-rating the quality of evidence, the domains for down-rating are more developed. For clarity, I will focus primarily on the moderating domains for down-rating the quality of evidence. Within these domains are many criteria, which pose *serious* or *very-serious limitations*, resulting in a down-rating by one level or two levels of quality, respectively. A brief description of these moderating domains can be found in Table 1.1.

Table 1.1: Moderating domains used to determine quality ratings on GRADE. Evaluative criteria are the elements of each domain that users assess the evidence by. Reasons for rating the quality of evidence down are given for each domain.

Moderating Domain	**Criteria**	**Rating**
Risk of Bias	Flaws in design or implementation of study, relative to ideal RCT methodology, e.g., lack of allocation concealment, lack of blinding, stopping early for benefit. Flaws assessed separately for individual studies, and for bodies of studies.	Each flaw in a study will mean a serious or very serious limitation and a consequent rating down by one or two levels.
Publication Bias	Suspicion that study results have been influenced by industry sponsors. Empirical examination should not be relied upon.	Suspect publication bias when evidence is from a number of small studies, especially if industry sponsored. Only rate down one level due to difficulty in determining actual malign influence.
Imprecision	3 criteria: 1 Confidence interval width 2 Optimal Information Size (OIS): calculation for what counts as an adequate sample size 3 Critical Thresholds: effect size thresholds of: a) no effect; b) important harm; c) important benefit	1 Is the confidence interval sufficiently narrow? If too wide, rate down one level. 2 If narrow, but OIS not met, rate down one level. If sample size very large, then do not rate down. 3 If OIS met, but CI includes ‘no effect’, rate down. 4 If 3) is met, but CI fails to exclude ‘important harm’ or ‘important benefit’, rate down by 1. If more than one of these steps is contravened rate down by 2 levels.

Continuation of Table 1.1

Moderating Domain	**Criteria**	**Rating**
Inconsistency	Heterogeneity of result across the studies in a body of evidence. 4 Criteria: 1 Similarity of point estimates; 2 Extent of overlap of confidence intervals; 3 Statistical tests for heterogeneity; 4 Explanations (population; interventions; outcomes; study methods) for heterogeneity if criteria 1-3 are contravened.	If large heterogeneity that is detected on basis of criteria remains after exploration of a priori explanations for that heterogeneity, rate down. If one criterion contravened, rate down one level; if more than one criterion is contravened, rate down two levels.

Indirectness	Differences between study and what we are interested in finding out (as defined by review question) in 3 criteria:
	1 Population;
	2 Intervention;
	3 Outcomes;
	A separate criterion is:
	4 Head to head comparisons between alternative interventions.
		Rate down if there are substantial differences for 1-3. What counts as substantial is dependent on context, but the difference must be likely to have an effect on outcome. Differences in one criterion results in rating down one level; differences in more than one criterion result in rating down two levels. For 4), if head to head comparisons are unavailable, but there is indirect evidence (e.g., both interventions separately against placebo) rate-down one level. If no comparisons direct or indirect available, rate down two levels.

Chapter 2

In defence of evidential pluralism

2.1 Introduction

Making a case in favour of evidential pluralism starts with identifying that evidence of correlation and evidence of mechanism have characteristic weaknesses, each of which are cancelled out by the strengths of the other kind of evidence. Recall that evidence of correlation between two variables A and B alone does not warrant an inference to the claim that A causes B. This is because there are explanations for the correlation other than that there is a causal relationship. Confounding is one explanation (§1.1). Another is temporal trends, where both variables increase or decrease over time for independent reasons. For example, bread prices and sea-levels in Venice have both risen over time, resulting in a correlation between them (Sober, 2001). But the rises are independent from one another: there is no causal link nor some confounder responsible for one or both of them. Another explanation is random error. Clinical trials are always at risk of this alternative explanation, but the risk is increased when the trial has an inadequate sample size. A result of random error is that any differences between trial groups can be attributed to coincidence, rather than a causal relationship. Statistical significance tests can help to indicate how probable it is that the results are due to random error. The p-value calculated by such a test is the probability that one would have observed a difference between trial groups, if it were the case that the difference was actually a result of random error. For a summary of other alternative explanations see Williamson (2019, p.4; Table 2).

Establishing a mechanism can help to rule out these alternative explanations. This is because it can explain how the putative cause brings about the effect. By ruling in this explanation, one rules out all other alternative explanations of the correlation other than that a causal relationship obtains. For example, finding a low p-value might not be enough to reject a random error explanation. Statistical significance tests come into a number of criticisms, including the problem of *p-hacking*, where researchers can test many possible associations in one test. The probability of finding one significant result increases as the number of associations tested increases (Stegenga, 2018, pp.157-158). Finding that a mechanism exists linking the putative cause and effect means one can rule out this *hacking* explanation. A similar process can be followed for ruling out alternative explanations that stem from confounding. Equally, evidence of mechanism alone cannot establish causation, due to problems of complexity and masking (see §1.1.2). Establishing a correlation helps to work out that there is a net effect, countering these problems. Therefore, only when one establishes a correlation and a mechanism can one establish causation. In sum, the evidential pluralism case against opposing views is that if one makes causal claims on either kind of evidence alone, then one does not rule out all alternative non-causal explanations.

Evidential pluralism has been criticised at length (Broadbent, 2011; Campaner, 2011; Howick, 2011a; Solomon, 2015, 2019). In this chapter I defend evidential pluralism against a number of different kinds of critiques. In §2.2 I respond to criticisms of EBM+, that if hold, would support the EBM view on evidence. I reject theoretical arguments that purport to show evidence of mechanism cannot be evidence for causation in §2.2.2. I argue that such claims are directed at mechanistic reasoning, whereas evidential pluralism requires carrying out a process of reinforced reasoning. In §2.2.3 I reject some counterexamples that purport to show that causation can be established on evidence of correlation alone. I appeal to an argument put forward by Williamson (2019), which claims that such counterexamples in fact only show that high-quality clinical studies can establish causation. This fact is compatible with evidential pluralism. I then consider a way in which a proponent of EBM can accept evidential pluralism but reject EBM+, namely, that evidence from mechanistic studies does not suffice to establish a mechanism. While in practice establishing a mechanism may be difficult, I argue that there is no *in principle* reason to think this. Finally, in §2.3 I respond to two different analyses of the case of the discovery that tobacco smoking causes lung cancer. If these analyses hold then a view somewhere in-between EBM and EBM+ would hold. My response to one analysis again relies on the disambiguation response, while my response to the other requires acknowledging that acting on the basis of evidence of causation does not imply that causation has been established.

2.2 Disambiguating criticisms of evidential pluralism

In this section I utilise the disambiguation approach to defend evidential pluralism against some common criticisms. The common theme to these criticisms is that if they hold, then a view on evidence more in line with EBM than EBM+ is defensible. Responses to them depend on acknowledging some distinctions between kinds of criticisms of evidential pluralism. *Theoretical arguments* appeal to general reasons for why evidence of mechanisms cannot play a role in evaluating causality. *Counterexamples* identify cases of scientific practice where in order to establish causation it was not necessary to establish both a mechanism and correlation. The two criticisms are linked. Theoretical arguments help to explain why the putative counterexamples hold. To see why this is the case, we first must distinguish between two kinds of putative counterexample, each with a different implication for evidential pluralism. The strongest kind shows that where evidence of correlation and evidence of mechanism *conflict*, the former trumps the latter. If such an example holds, then evidential superiority holds, and EBM is a more complete epistemology of causality in medicine than EBM+. The other, weaker kind of putative counterexample shows that evidence of correlation alone is sufficient to establish causation. One might think this kind of putative counterexample shows that evidential monism is true. However, it is also consistent with a broader evidential diversity thesis. To make an inference from a *correlational counterexample* to an evidential monist conclusion one also has to level a theoretical argument. The purpose of such an argument is to justify why this kind of counterexample holds more generally, such that evidence of mechanism cannot provide evidence of causation. Additionally, in some cases, one can interpret examples in a way that favours both evidential pluralism and a rejection of it. Counterexamples are thus often of limited use. In such cases theoretical arguments are again levelled to support interpretations that lead to rejection of evidential pluralism.

2.2.1 Counterexamples

One can differentiate even further among the kinds of ‘conflict’ counterexample. The strongest kind of counterexample is where evidence of correlation conflicts with evidence of mechanism such that each line of evidence supports a different conclusion, e.g., one supports causation, one does not. Such a counterexample purports to reject evidential pluralism, as the conclusion made in that case is warranted on the basis of the evidence of correlation alone. One way this happens is for evidence of mechanism to support a causal claim, but evidence of correlation to later overturn the causal conclusion. Call such cases *overturning* examples. Howick (2011b, pp.154-156) cites 20 examples of conflict where evidence from comparative clinical studies overturned evidence from mechanistic studies. Another way is for one line of evidence to establish a correlation while another line finds no support for the existence of a mechanism, yet causation is still established. Call such cases *defeating* examples, one of which is the Semmelweis case (see below for details).

There are many counterexamples that purport to show that causation can be established on the basis of evidence of correlation alone. Solomon (2015, p.128) argues that effective treatments for cystic fibrosis are supported solely by evidence obtained from high-quality RCTs. Moreover, the mechanisms linking these interventions and beneficial outcomes for cystic fibrosis patients are not understood. So, evidence from clinical studies is all that was needed to establish the effectiveness of these treatments. Howick (2011a) has a list of treatments where he thinks effectiveness was established solely on evidence from comparative clinical studies, and where no mechanism was established. Examples include aspirin for pain relief and deep brain stimulation (DBS) for Parkinson's disease. Plutynski (2018) and Gillies (2019a) separately argue that evidence of mechanism was not needed to establish that smoking causes lung cancer. In that case, evidence of correlation alone sufficed.

One way to respond to putative counterexamples is to reject their interpretation, in that one can cast the example as also supporting evidential pluralism. Solomon is not clear on what extent of understanding is needed for a mechanism to be established. For example, she argues that "no one understands yet why hypertonic saline mist works better than water mist, or why ibuprofen works better than corticosteroid" (Solomon, 2015, p.128) as treatments for cystic fibrosis. Elsewhere she notes that "although our knowledge of some basic mechanisms in cystic fibrosis has grown, we do not understand these mechanisms fully" (Solomon, 2015, p.130), indicating that she requires establishing the details of an exact mechanism, rather than the existence of a mechanism. And all that evidential pluralism requires is that the existence of a mechanism is established, not its exact details (Williamson, 2019). Gillies (2019a, pp.177-181) develops this response and levels it against some of Howick's putative counterexamples. Gillies argues that it is always possible to develop a mechanism in more detail, but not always necessary. With regards to aspirin and DBS, there is evidence of a mechanism but not of the exact mechanism. But there is enough evidence to confirm that a mechanism does exist, so evidential pluralism is satisfied. Gillies deals with most of Howick's remaining putative counterexamples by showing that they are 'anachronistic', because they appeal to cases that occurred before the statistical techniques employed in comparative clinical studies were developed. Gillies pegs this date to around the mid-19th Century. An example of Howick's that is anachronistic is Edward Jenner's introduction of the smallpox vaccine before it was understood how the vaccine works. The main problem with this kind of response is that it hinges on vague notions such as 'extent of understanding' and the 'date at which the right statistical techniques were introduced'. One can again interpret such notions in ways that lead to conclusions in favour of either evidential pluralism or its negation. Another way to resolve the issue is needed.

2.2.2 Theoretical arguments

To bolster their conclusions against responses made to the interpretation of putative counterexamples, critics of evidential pluralism make theoretical arguments. By rejecting evidential pluralism on general terms, there is no need to resort to disagreement over vague notions. A number of such arguments that reject evidential pluralism take aim at the problems of complexity and incompleteness facing *mechanistic reasoning* (Howick, 2011b; Solomon, 2015), where mechanistic reasoning "is an inferential chain (or web) linking the intervention...with a patient-relevant outcome, via relevant mechanisms." (Howick, 2011a, p.930). Many mechanisms of interest in medicine are complex, in the sense that they consist of a large number of components with an often non-linear organisation. This complexity raises two interlinked problems for the epistemology of mechanisms. One is that our evidence of mechanism, or our mechanism description based on that evidence, will almost always be incomplete. Moreover, if we think we have a complete mechanism, it is still possible that we are missing evidence for components of it. The incompleteness problem is linked to a predictability problem: when mechanisms are complex, due to the likely incompleteness of our knowledge of them, the effects of such mechanisms are impossible, or at least very difficult, to predict. A related problem is that any one intervention may initiate more than one mechanism, so evidence that there are mechanisms linking the intervention and outcome is not enough for predictive purposes. This is because the effect of other mechanisms may outweigh the effect of the intervention mechanism, nullifying its *net effect*. This kind of argument can explain why it is the case that actual instances of scientific practice do not adhere to the stipulations of evidential pluralism. If one cannot predict the effects of mechanisms then this is why evidence of mechanism should not be used to support causal claims.

The evidential pluralism response is of course that both evidence of correlation and evidence of mechanism are required. Evidential pluralists can accept that it is rarely if ever possible to infer that an intervention has a beneficial effect from evidence that a mechanism operates. The pluralist instead requires obtaining evidence of correlation as well, which provides evidence of a net effect. The arguments levelled against mechanistic reasoning are levelled in fact against *standalone* mechanistic reasoning. Whereas the evidential pluralist requires reinforcing mechanistic reasoning with a piece of correlational reasoning, where correlational reasoning involves inferring from evidence of a correlation between an intervention and outcome, that the intervention causes the outcome (Auker-Howlett and Wilde, 2020). Moreover, correlational reasoning alone is not enough to infer causation, due to the problems facing evidence of correlation. A process of *reinforced reasoning* infers causation from evidence that establishes a correlation and a mechanism because each kind of standalone reasoning helps to rule out the possible errors the other kind faces. This is the driving motivation for evidential pluralism and why arguments against standalone mechanistic reasoning do not work. Therefore, such criticisms also do not work against evidential pluralism.

This response also deals with 'overturning' counterexamples. If we interpret those counterexamples as saying that clinical studies found that there was no evidence of a correlation, and this overturned evidence of mechanism that supported a causal claim, such a case is still consistent with evidential pluralism. This is because we cannot establish causal claims on the basis of evidence of mechanism alone for the exact reason that it is difficult to establish a net effect on the existence of a mechanism alone. Finding no net effect and not concluding causation is entirely consistent with evidential pluralism, as establishing causal claims requires establishing both a correlation and a mechanism. If a process of reinforced reasoning was carried out, causality would not have been supported in the first place, hence no evidence of mechanism to overturn.

A critic of evidential pluralism may object directly to the reinforced reasoning argument. Because reinforced reasoning resolves issues directed at standalone mechanistic reasoning, this objection would have to deny the ability of an established mechanism to rule out the chance of confounding. It is difficult to see how one could make this argument. Why this argument is difficult to make can be clarified by viewing the process of using an established mechanism to rule out confounding as *just one method* for inferring causation from evidence that establishes that a correlation exists. What this proponent of EBM would ignore are the commonalities between randomisation and 'establishing a mechanism' as methods for ruling out confounding. To illustrate the commonalities, consider what is said about randomisation by its critics.

Criticisms of the view that randomised studies are better than non-randomised studies often focus on whether randomisation can rule out confounding better than some other method. For example, recall Worrall (2002)'s argument (§1.1) for why the method of matching known confounders across treatment and control groups in non-randomised studies is no worse a method for ruling out confounders than randomisation. It is typical of the fact that debates over the supremacy of randomisation do not reject randomisation as *a method* for inferring causation from an observed correlation. But they do support the view that randomisation is just one tool for ruling out confounding, even if it can be a particularly powerful tool. Non-randomised studies have similar tools at their disposal. The analysis of the non-randomised studies that found an association between saturated fat and heart disease of Gillies (2019a, pp.115-118) identifies a number of such tools. For example, researchers had identified a difference in cardiovascular disease between one population from Japan and one from U.S.A., which they attributed to the lower amount of saturated fat in the Japanese diet. A plausible alternative explanation was that Japanese people had genes that were protective against cardiovascular disease. This explanation was ruled out by comparing groups of Japanese who lived in the U.S.A., and were eating a typical American diet, with Japanese living in Japan, who were eating a Japanese diet. Death by cardiovascular disease was higher in the American diet cohort, thus ruling out the genetic alternative explanation.

At root, there is nothing special about a method for inferring causation other than its ability to rule out alternative non-causal explanations of a correlation. Establishing a mechanism linking the putative cause and effect is also a method for ruling out alternative explanations. It does so by the process outlined in §2.1. So the position that establishing a mechanism cannot be used to rule out confounding would not work.

2.2.3 Clinical studies can alone establish causation

Recognising that one must do a process of reinforced reasoning does not help to reject counterexamples that purport to show causation can be established on evidence of correlation alone. Reinforced reasoning is rejected if those counterexamples hold, as only correlational reasoning was carried out. To reject such putative counterexamples, proponents of evidential pluralism can instead advance their own theoretical arguments, which go beyond the arguments made for the original thesis. Indeed, Williamson (2019) accepts the general principle that evidence from *clinical studies alone* can establish causality, but argues that taking a disambiguated view means this fact supports evidential pluralism rather than refuting it. When Solomon and Howick both argue that evidence of correlation can alone establish causation, the counter-examples they level in support instead show that evidence from *high-quality clinical studies* alone can establish causation. But, given Illari's distinction, it is possible that clinical studies can provide both evidence of correlation and evidence of mechanism.

Williamson argues that sufficiently high-quality clinical studies can establish a mechanism when "the threshold is reached for establishing a genuine correlation, and that bias and confounding are ruled out as explanations of this correlation." (2019, p.44). Bias and confounding are ruled out when there are a large number of high quality clinical studies pointing in the same direction that together observe a large correlation. Recall from §1.1 that the supremacy of randomisation has been challenged on the grounds that there is a chance that in any one randomisation, unknown confounders are not balanced across trial groups. By requiring a large number of studies, this problem is mitigated. In addition, one must also rule out other explanations of the correlation, such as temporal trends, and mathematical or logical relationships. As long as there is no evidence against the existence of a mechanism (e.g., the mechanism would conflict with established theory), then if all these conditions are fulfilled there must be a causal explanation of the correlation. Therefore, there must be a mechanism linking the putative cause and effect that explains the correlation.

One might object to the assumption that when there is a causal explanation it is a mechanism that explains the correlation. One way is to object to whether mechanisms are how science causally explains. For example, Dupré (2013) argues that it is processes, not mechanisms, that are fundamental to causation in biology. While the distinction between the two is subtle, the key difference can be put down to the emphasis each account puts on 'change'. Mechanisms can be conceived of as productive of regular input-output links consisting of stable entities. Processes on the other hand can be conceived of as a flow of physical material that only briefly exhibits stability. One response to this dispute is that it is metaphysical in nature, rather than epistemological. It could be the case that, at root, life is processural, rather than mechanistic. But at least in medical and biological practice, when seeking to explain how an effect is brought about by a putative cause, scientists talk of mechanisms. Dupré contests this point, and suggests that those scientists are mistaken, and cites other scientists that talk in terms of processes. In lieu of a final judgement on the metaphysics of causation in biology, what we are interested in here is what link obtains between two variables, such that we can causally explain a correlation. The basic point is that there must be some sequence of steps linking the putative cause and effect. The characterisation of mechanisms utilised in evidential pluralism, and in this thesis, is one way of characterising this link. So to say that there must be a mechanism is not wrong even if it turns out that biology is fundamentally processural.

The only other option left open to the opponent of evidential pluralism would be to argue that even when there is a causal explanation of a correlation between two variables, that no mechanism exists linking them. This option would be unpalatable even to them, as they are committed to the existence of *some* mechanism linking the variables. When one accepts that there has to be some sequence of steps that links two correlated variables, and that high-quality clinical studies rule in a causal explanation, then such studies also have to suffice to establish *both* a correlation and a mechanism. Therefore, the counterexamples that purport to show that evidence of mechanism is not needed, only show that evidence from *mechanistic studies* was not needed.

This line of argument may also make sense of the final *defeating* conflict counterexample: the Semmelweis case. This putative counterexample is especially tricky as proponents of evidential pluralism also cite it as support (Russo and Williamson, 2007, p.163). The case is well known and is treated at length in Gillies (2005), but the response will need some explanation of the details of the case. Briefly, Semmelweis investigated the cause of high rates of puerperal fever in a Viennese maternity clinic. He conducted what would now be called comparative clinical studies, and found that a correlation existed between clinical staff not washing their hands after carrying out autopsies (W) and high rates of puerperal fever in women examined by those clinical staff (P). Moreover, Semmelweis conducted his investigations in such a way that he ruled out the plausible alternative explanations to his own, namely, that W caused P. His conclusion was however not accepted at the time as it ran counter to the dominant theory of disease transmission, namely, *miasma theory*, where disease is transmitted by 'bad air' that smelt foul and resulted from decaying organic matter. On this theory, unwashed hands could not be a cause. Semmelweis's conclusions were only supported by the scientific community once the germ theory of disease had been developed. This allowed a mechanism to be established explaining the correlation between W and P in terms of micro-organisms being transferred from unwashed hands to women in childbirth. So evidential pluralists level this example as support. But critics cite this as a rejection of evidential pluralism as a normative thesis: clinicians should not have followed evidential pluralism, as if they had accepted the conclusions of the clinical studies, then further deaths would have been prevented (Broadbent, 2011; Howick, 2011a).

The line of argument developed in this subsection may support the evidential pluralist's interpretation. Semmelweis's investigations should have ruled in the existence of a mechanism linking W and P, as he ruled out all plausible alternative explanations of the correlation other than a causal explanation. If the comparative clinical studies are as good as EBM proponents say they are, then the studies should have been good enough to rule in the mechanism as well. However, one might object to this analysis and say that the miasma theory, the dominant paradigm for disease transmission of the day, had already ruled out the existence of a mechanism linking W and P. In other words, the dominant theory by which one would explain the correlation conflicts with the evidence of correlation. This would mean the disambiguated approach fails to reject the criticism of evidential pluralism as it is not the case that a mechanism was also established; at best, the existence of a mechanism was doubtful. Causation was instead established on evidence of correlation alone, in the face of evidence of mechanism that pointed in the other direction.

One response to this analysis is made by Gillies (2019a). Taking a Kuhnian approach, he argues that this kind of conflict will only happen during a scientific revolution, and as such revolutions are rare, evidential pluralism will hold in almost all cases. However, this response does not support evidential pluralism, but rather a broader evidential diversity view. I think that the disambiguated approach does work. This is because if high-quality comparative clinical studies provide evidence that runs counter to biological theory, then this gives reason to doubt the theory. There will also be reasons to accept the theory, whether that be evidence from mechanistic studies, clinical studies or background knowledge. An evaluation of the evidence for both a correlation and mechanism should include all pertinent evidence as support for both objects. Such an evaluation, if a case such as Semmelweis's occurred today, should not lead to rejecting causality due to conflict with currently accepted mechanistic theory. Instead, the evidence from clinical studies should outweigh the evidence of mechanism from other sources. In Semmelweis's time, mechanistic studies did not have the experimental rigour, nor was the background knowledge as comprehensive, as it is today. So it is more likely that good comparative clinical study evidence would defeat the accepted theory of that time, if one were to take an evidentially pluralist approach. The basic point is that clinical studies are one method for obtaining evidence of mechanism, and so should count as part of evaluating the total evidence for some mechanistic theory. They may, but not always will, defeat evidence from mechanistic studies.

2.2.4 EBM+

A proponent of EBM may accept all the arguments made in this section, and still object to EBM+, rather than evidential pluralism *per se.* Note that EBM+ is motivated by, but not equivalent to evidential pluralism, which makes no stipulation on what methods can be used to obtain evidence of correlation and mechanism. One might object to the ability of mechanistic studies to establish a mechanism, but accept the disambiguated EBM where high quality clinical studies suffice to establish both a correlation and a mechanism. This kind of EBM would keep to the letter, but not the spirit, of evidential pluralism. Moreover, the practice of evaluating causality in medicine would not change, removing the force of evidential pluralism by removing a significant implication of it.

Firstly, consider a case where the proponents of EBM and EBM+ would disagree with respect to whether a causal claim is established, while at the same time agreeing that high-quality clinical studies may suffice to establish both a correlation and a mechanism. This is the case used to support reinforced reasoning by Auker-Howlett and Wilde (2020), namely, establishing the effectiveness of a pegylated version of interferon (peg-interferon) and ribavirin as a treatment for Hepatitis C. More precisely, the conclusion that peg-interferon and ribavirin is a more effective treatment than standard interferon and ribavirin is only warranted by the evidential pluralism position. This is because evidence from clinical studies sufficed to establish a correlation between treatment and suppression of Hepatitis C viral load, but could not rule out *all* sources of confounding (rather than just ruling out plausible confounders). Improperly implemented randomisation and absence of blinding prevented inference to causation from clinical studies alone. However, a mechanism that can explain why peg-interferon is more effective for treating Hepatitis C than standard interferon was established by evidence obtained from mechanistic studies. Mechanistic reasoning would not suffice to establish causation, as it was likely that there were other mechanisms operating that could cancel out the effect. So only the EBM+ approach can make sense of why peg-interferon was considered more effective. But as Williamson (2020) notes, it is not clear that decision makers did explicitly follow this kind of reasoning. An EBM proponent could respond that while the treatment was recommended, strictly, it should not have been. This is a point at which a proponent of EBM who accepts that high-quality clinical studies suffice to establish both a correlation and a mechanism would disagree with the proponent of EBM+. To do so they would have to object to establishing a mechanism by means of mechanistic studies.

One might attempt to support this view by claiming that due to practical concerns one cannot establish a mechanism on the basis of mechanistic studies. A potential way to make this argument is to invoke the complexity and incompleteness worries that were levelled against standalone mechanistic reasoning. Such worries do not pose a problem for predicting the effects of a mechanism when one takes an evidentially pluralist view. But one may still worry that they pose a problem for establishing that 'a mechanism exists on the basis of evidence from mechanistic studies'. Solomon (2015) seems to argue for this point as well. She argues that while we may require that the complexity and incompleteness of mechanisms should be acknowledged when using them to infer causation, "it is difficult to assess our knowledge of the completeness of relevant mechanisms" (2015, p.122-123). Completeness is a worry for evidence obtained from mechanistic studies as evidence from each study is typically of one or a few components. The evidence for the mechanism as a whole is then pieced together from many studies.

However, this is not an *in principle* argument against establishing a mechanism on the basis of evidence from mechanistic studies. This is a worry about *assessing* completeness, whereas the arguments considered here are about whether it is possible in principle to use evidence of mechanisms to establish causation. Whether our evidence of mechanism is complete is a question for the *evaluation* of evidence from mechanistic studies. I will address this question when I address practical concerns about evaluating evidence from mechanistic studies in Chapter 4. But questions about evaluation should come after we decide what kinds of evidence, in principle, we should admit to the evaluation. It is not clear whether an argument can be made that in principle, establishing a mechanism on the basis of evidence from mechanistic studies is no worse than on the basis of clinical studies.

The disambiguation approach, and the reinforced reasoning and 'clinical studies alone' arguments deal with most putative counterexamples. They at least deal with those that would support an EBM view on evidence. They may not however deal with a putative counterexample that would support a view on evidence that fell somewhere between EBM and EBM+. The example was introduced at the start of this section: the case of whether tobacco smoking causes cancer (Plutynski, 2018; Gillies, 2019a). I turn to this case next.

2.3 Tobacco case

The *tobacco case*, analysed separately by Plutynski (2018) and Gillies (2019a), is a counterexample in a similar vein to those posed by Solomon (2015) and Howick (2011a): evidence of correlation alone sufficed to establish causation. However, neither author uses it to argue for an evidential monism or superiority view, although both do reject evidential pluralism. Instead, a broader evidential diversity view is maintained, and this view on evidence falls somewhere in between EBM and EBM+. If either analysis holds, then EBM+ is not a more complete epistemology than EBM on the question of what counts as the best evidence for causality in medicine. Importantly, their criticisms of evidential pluralism are not easily susceptible to responses that utilise the disambiguation approach. To be precise, Gillies's analysis is made within a wider philosophical framework that at first sight does not support Illari's distinction. There is however a way to make his view amenable to the disambiguation approach. Plutynski on the other hand holds a view that is not amenable to the disambiguation approach. I thus offer a novel response to her position. The details of the tobacco case are central to the way each author makes their arguments, so I will set them out first. I will then address in turn each argument made against evidential pluralism.

It is now established that tobacco smoking is a cause of lung cancer. It is the contention of Plutynski and Gillies that this causal claim was established before a mechanism linking tobacco smoking and lung cancer was established, contra evidential pluralism. The earliest evidence for the existence of a correlation between smoking and lung cancer was obtained in retrospective cohort studies in the 1950s. Those studies established that there was a correlation in the underlying probability distribution. The question was whether this correlation had a causal explanation. As the study design was non-randomised and observational, specifically of a cohort study design (see description of clinical studies in §1.1), the researchers could not rule out many kinds of bias, including interviewer bias, selection bias, and confounding. In addition, these cohort studies were retrospective, meaning the data was obtained from existing records. In this case, the exposure data consisted of self-reports of past events, namely the smoking habits of participants. This introduced an additional bias, namely *recall bias*, where results can be biased by the unreliability of participants' memory of events. To conclude that this correlation was causal involved showing that the potential biases could not account for the correlation and its size. Techniques other than randomisation were used for ruling out these alternative explanations. Doll and Hill considered plausible confounders and argued that they did not account for the correlation. Critics, such as the influential statistician R.A. Fisher, argued that they did. For example, he postulated a genetic confounder, where a gene existed that accounted for a pre-disposition to both smoke and to get cancer.

Prospective cohort studies were then carried out, where a study is prospective when it follows the participants over some period of time. For example, Doll and Hill asked doctors to describe their smoking habits, before following up on the health of those doctors. From 40,564 doctors interviewed, 35 died from cancer, all of whom smoked. Moreover, those who died smoked more on average than those who did not die. In favour of these results was that prospective studies do not depend on participants recalling smoking habits at a time in the past, which mitigates recall bias. The observation of a dose-response curve (the greater the amount of smoking, the likelier it was that the person got lung cancer), was also evidence that confounding is not an explanation of the correlation. While there were still critics, the surgeon general of the U.S.A. published a report in 1964 reviewing 6,000 articles in 12,000 journals and concluded that there was a correlation between tobacco smoking and cancer. Smoking use declined from then on. Further details of the case can be found in the relevant chapters of their books (Chapter 5 for Plutynski, Chapter 8 for Gillies).

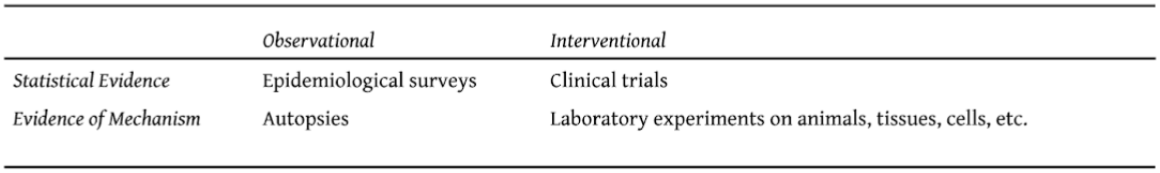
2.3.1 Gillies

In 1976, Doll and Peto reported a follow up on the prospective cohort study of doctors' smoking habits, which showed that:

> "smokers are on average more than 10 times more likely to die of lung cancer than non-smokers and this figure rises to more than 22 times for heavy smokers who consume 25gms or more of tobacco every day. The results [were] highly statistically significant." (Gillies, 2019a, p.137)

Gillies argues that this was the latest point at which the causal claim 'tobacco smoking causes lung cancer' could have been established, and in the paper Doll and Peto did indeed conclude the association was causal. Moreover, Gillies claims that it was not until the 1980s that the mechanism linking tobacco smoking and cancer was established. He takes this as a rejection of evidential pluralism, as the causal claim was established without establishing a mechanism.

For Gillies, causation was confirmed because Doll and Peto had ruled out all possible alternative explanations of the established correlation. And I have argued that when this occurs a mechanism is also established. However, Gillies does not think that this alone was enough to establish causation. Why he thinks this, depends on the metaphysical theory of causality he defends earlier in his book: the action-related theory of causality, where an action-related theory "stresses the connection between causal laws and interventions" (Auker-Howlett and Wilde, 2019, p.387). A consequence of holding this theory is that causation can only be confirmed when one has *interventional* evidence, where interventional evidence is obtained through experimental methods, e.g., RCTs, mechanistic studies (§1.1). Instead of maintaining Illari's distinction, he instead holds a two-by-two distinction: one between statistical evidence and mechanistic evidence; another between observational and interventional evidence. It is two-by-two as one can have observational statistical or



	Observational	*Interventional*
Statistical Evidence	Epidemiological surveys	Clinical trials
Evidence of Mechanism	Autopsies	Laboratory experiments on animals, tissues, cells, etc.

Figure 2.1: Table showing Gillies's two by two classification of types of evidence in medicine (Auker-Howlett and Wilde, 2019).

mechanistic evidence and interventional statistical or mechanistic evidence (see fig. 2.1). Because of his theory of causality, only by intervening on potential causes can we confirm causation. The problem Gillies identifies is that the epidemiological studies by which the evidence for the causal claim was obtained were solely observational. This is often the case where the intervention would be harmful to participants. In this case intervening to make one cohort smoke, while the other cohort did not, would be highly unethical.

Because of this principle of interventional evidence, his analysis of the tobacco case leads him to conclude that one does not need to *establish* a mechanism. Instead, one need only show that a mechanism is plausible, as long as the evidence for that plausible mechanism comes from interventional means. In the tobacco case, background knowledge included evidence that some chemicals could cause cancer, and that tobacco smoke contained a large number and variety of chemicals. While steps by which the exact chemicals in tobacco smoke led to lung cancer were not known, it was plausible on the background knowledge about the causes of cancer that there was a mechanism linking smoking and lung cancer. And this evidence was obtained in mechanistic studies, so the principle of interventional evidence was fulfilled. His analysis of the tobacco case leads him to make a qualification to evidential pluralism, namely that to establish causation one need not establish a mechanism, but merely show that a mechanism is plausible, where a plausible mechanism is one supported by background knowledge. In this sense he is committed to a broader evidential diversity view, as he does not think *establishing* a mechanism is necessary for establishing causation.

One response to Gillies might put the disagreement down to a difference in metaphysics.

Gillies's analysis of the case is constrained by his commitment to an action-related theory of causality. In the same paper that introduced RWT, Russo and Williamson use the thesis to argue for a competing metaphysical theory of causality, namely, the epistemic theory of causality (Russo and Williamson, 2007). Evidential pluralism is therefore not held to the principle of interventional evidence. To argue that Gillies is wrong in his analysis of the tobacco case and evidential pluralism, one may have to refute his action-related theory of causality. But this sort of resolution is likely not forthcoming. Indeed, part of the motivation for evidential pluralism, and the epistemic theory of causality, was that analyses of causality that focused on metaphysics were unable to convince one another, and the debates were at an impasse. Instead, what counts in medicine is the epistemology of causality, so moving away from metaphysics would advance debates in this area. However, Gillies might not accept this as he links the epistemology of causality to metaphysics.

There is a way to respond that does not avoid the metaphysical issue. Auker-Howlett and Wilde (2019) argue that the reasoning involved in the disambiguation response is also able to satisfy the principle of interventional evidence. This is because Gillies defines mechanisms as sequences of causes, and causes are defined in terms of his action-related theory of causality. He also allows that one can obtain observational evidence of a mechanism (see fig. 2.1), which implies that one can get evidence by observational means that satisfies the principle of interventional evidence. This kind of response only makes sense when one applies Illari's distinction, which Gillies does not always do. In this case, the object of evidence is interventional in nature, whereas the method by which the evidence is obtained is observational. This analysis carries over to evidence of mechanism obtained by means of observational clinical studies. Given that mechanisms are defined in terms of the action-related theory of causality, the objects of this evidence also satisfy the principle of interventional evidence. So the reasoning involved in establishing a mechanism linking tobacco smoking and lung cancer from evidence obtained by means of observational clinical studies also satisfies the principle of interventional evidence. Gillies's analysis of the tobacco case is thus compatible with evidential pluralism.

2.3.2 Plutynski

Plutysnki rejects evidential pluralism, but is in favour of evidential diversity. Her analysis of the tobacco case leads her to reject evidential pluralism on the basis that evidence of mechanisms was not necessary to confirm causation. However, her analysis of another case study, the Downwinders case (see Plutynski (2018, pp.148-155)), leads her to conclude that mechanistic evidence can be used to establish causal claims alongside evidence from clinical studies. Note that her view is more in line with standard EBM than Gillies, as she also thinks that clinical studies alone suffice to establish causation. Importantly, the responses outlined above to putative correlational counterexamples may not work against her analysis. I first outline her arguments.

Plutynski argues that "long before a mechanism was established, warning the public of the risk of tobacco was (arguably) warranted", going so far as to state that "in 1965, in Hill's view, the evidence was more than sufficient to warrant public policies warning smokers of risks to their health" (2018, p.145). Moreover, she agrees with Gillies that a mechanism was not established until the 1980s. Because of this, in Chapter 5 of her book, Plutynski claims that evidential pluralism is false. To do this she argues for three separate claims: causation is never established, where established is understood as certainty; that there are no necessary and sufficient conditions for causality; making a judgement of causality involves ruling out alternative explanations. The first and third claim do not refute, and are consistent with, evidential pluralism. Under RWT, establishing is fallible, and does not entail absolute certainty. Instead, establishing means that "standards are met for treating the claim itself as evidence, to be used to help evaluate further claims" (Williamson, 2019, p.35). Williamson also notes that it is unclear whether establishing is factive, as it is possible for a previously established causal claim to be defeated. Establishing is still however a "high epistemological standard....and should be distinguished from acting...as a precautionary measure" (Williamson, 2019, p.36). With respect to Plutynski's third claim, ruling out alternative explanations as a means to establishing causation is both a major contention of this thesis, and central to the justification of evidential pluralism. Therefore, only Plutynski's second claim, that there are no necessary and sufficient conditions for causation, is a potential refutation of evidential pluralism.

One problem with her argument is that it is aimed at mechanistic evidence, rather than evidence of mechanism. While she does not state a precise definition, it is clear from what she requires for establishing a mechanism that mechanistic evidence is both evidence from mechanistic studies, and evidence for the details of an exact mechanism. Because of this, Plutynski's rejection of evidential pluralism could fail on two accounts. Firstly, evidential pluralism does not require the details of an exact mechanism. As Plutynski admits, mechanistic studies did indicate a partial mechanism around the same time as the epidemiological studies identified the correlation. So, using the absence of the details of a mechanism to refute RWT could fail as that is not what is required to establish causality. The point at which a mechanism was established is however agreed on by both Gillies and Plutynski, and Gillies does not require the details of a mechanism for it to be established. So, it is equally plausible that evidence from mechanistic studies did not establish a mechanism until the 1980s. I concede this point and argue against their analysis of when a mechanism was established. Secondly, one can take the disambiguation approach levelled at Gillies: it is plausible that eliminating alternative explanations also ruled in the existence of a mechanism. This interpretation would be in line with evidential pluralism.

A response that relies on the disambiguation approach, as mine does here, may not work for Plutynski. This is because of what she requires for confirming causation. Plutynski argues that we should be mindful of the distinctions between belief, evidence, and action. What evidence we have and what we believe on the basis of that evidence should be independent of how we act. What counts as sufficient evidence for action depends on how we are putting that evidence to use. For Plutynski, at issue is not "do we know this causes that?" but "do we have enough evidence to warrant this or that action" (Plutynski, 2018, p.128). How strongly our evidence is in favour of some hypothesis should lead us to hold equally strong beliefs, but how we act will depend on context. Some of our beliefs may be about what we should do in non-risky situations, so we may require less strong evidence than we would in more risky situations. She argues that causation is a matter of (probable) degree, and that the available evidence in the tobacco case gave warrant to act to prevent smoking. It is consistent with this view that epidemiological studies did not establish a mechanism as well as a correlation, contra evidential pluralism. Moreover, evidence from mechanistic studies did not suffice to establish a mechanism until the mid-1980s, and she argues that confirming the causal link was warranted long before this. So, evidential pluralism is false as evidence of mechanism was not necessary to establish the causal claim.

The surgeon general's report did indeed confirm that there was a genuine association, a conclusion consistent with the idea that a correlation was established. But this does not mean that causation was established as well. It would be a mistake, so this response goes, to infer that causation was confirmed on the basis that public health officials moved to act to prevent smoking. One can make the case, using her own analysis, that public health officials were warranted to act to prevent smoking, even if causation was not strictly established. The causal claim may have been highly probable given that there were few possible alternative explanations not ruled out. But given that not acting would likely contribute to excess deaths due to lung cancer, it was legitimate to act to prevent smoking. The mistake Plutynski makes is to reason from the conclusion that one should act to stop smoking, to the conclusion that the causal claim was established. In other words, action does not imply establishing, while establishing plausibly does imply acting.

One way to think of this is that whether we decide to act may depend on whether our beliefs cross some pre-defined threshold (Williamson, 2010, p.66-70). This idea relies on the notion that beliefs come in degrees (an idea developed in more detail in Chapter 7), an assumption that is harmless for present concerns as the dispute is over whether making causation highly probable is enough to establish it. Thresholds may be set relative to the risk involved in acting or not acting. In the case of preventing lung cancer, where the risk of not acting if the claim turns out to be true is high, our thresholds for acting may be lower than for an action for which we require high thresholds. For example, we may require a high degree of belief that the liquid in an unmarked jar with skull and crossbones on it is not poison before we decide to drink it. And it seems plausible that deciding to act to prevent smoking requires a threshold for belief lower than what it would take to establish a causal claim, which is a high epistemological standard. Evidential pluralism is an epistemological thesis about what it takes to establish a causal claim, but says nothing about what it takes for someone to act on one's beliefs. So an analysis of the tobacco case that says the evidence was sufficient to warrant action, even if the evidence does not establish a mechanism, is not an argument against evidential pluralism.

A higher standard is arguably worth holding, as the point of evidential pluralism is to evaluate what causal relations hold in general, rather than just in a decision making context. Consider the full quote from (Williamson, 2019, p.36), that sets out what it takes to establish:

> "Whether or not establishing is factive, it requires meeting a high epistemological standard. In particular, establishing a causal claim should be distinguished from acting in accord with a causal claim as a precautionary measure: in certain cases in which a proposed health action has a relatively low cost, or failing to treat has a high cost, it may be appropriate to initiate the action even when its effectiveness has not been established, so that benefits can be reaped in case it turns out to be effective."

This approach to establishing puts the truth of claims at a paramount rather than considerations about what actions are right to take. Moreover, there is plausibly a place for both in a more general epistemology of *medicine*, but not in an epistemology of *causal evaluation* in medicine. Moreover, Plutynski's view is ultimately compatible with EBM+. In the next Chapter, I introduce the evidence evaluation framework built on EBM+ principles, and central to it is a graded notion of causation. Such a graded notion allows a more fine grained evaluation of causality, and is central to all evaluative frameworks. Integration into a decision making context would also benefit from such a graded notion. But this should not be mistaken for what it takes to establish causation.

2.4 Conclusion

I have argued that when one takes a disambiguated approach, criticisms of evidential pluralism can be rejected. In particular, arguments against mechanistic reasoning, or counter-examples that purport to show causation can be established on evidence of correlation alone, fail when we disambiguate. Evidential pluralism both requires carrying out a process of reinforced reasoning, and can accommodate the fact that high-quality clinical studies sometimes suffice to establish causation. Given a disambiguated approach, critics of evidential pluralism have few resources to fall back on. They could argue that causation can be established without ruling out all alternative explanations. But this contradicts a basic principle that evidential monists use to justify EBM. It is also implausible that ruling out all alternative explanations does not rule in a mechanism, given the importance of mechanisms as causal explanations in medicine, a point which EBM proponents concede. I thus posed a way that EBM could maintain the supremacy of clinical studies which cannot be rejected by purely taking the disambiguated approach. I argued in response that there is no *in principle* way to reject establishing mechanisms by means of mechanistic studies. Moreover, when we see establishing a mechanism on the basis of mechanistic studies as just one method for ruling out confounding, then we see how the EBM position cannot maintain the supremacy of clinical studies. The final refuge would be to come up with some alternative account of what counts for confirming causation. The account proposed was one based on the idea that what evidence is sufficient for confirming causation is evidence that warrants action. This would however likely be a lower epistemological standard than the characterisation of ‘establishing’ employed in EBM+. This lower standard may be appropriate in a decision making context, where considerations of risk must also be made. But it is not appropriate when we are seeking to work out what causal relations hold in general. So a methodology for causal evaluation that is evidentially pluralist is better than one that is evidentially monist. On this count, EBM+ is a more complete epistemology of causation than EBM.

Part II

Practice and concepts

Chapter 3

Feasibility

3.1 Introduction

I have defended EBM+ on the grounds that it offers a better account of evidence for causation. Specifically, an epistemology of causation that is evidentially pluralist is better than one that is not. EBM is however not just concerned with what kinds of evidence are admissible. It also makes prescriptions on the *evaluation* of evidence. Indeed, its original mission statement was that “Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence” (Sackett et al., 1996, p.71). To this end, numerous frameworks have been developed in the EBM tradition for explicitly evaluating evidence from clinical studies. An evaluation of evidence ultimately considers whether an evidence base consisting of clinical studies that purports to have established causation, does in fact do so. Evidence bases are evaluated as a whole, as not all studies will be positive, and one must take into account both positive and negative results. To be considered a more complete epistemology of causality in medicine, EBM+ must also be able to accommodate the evaluation of evidence. Evidential pluralism stipulates what it takes to establish causality, but says nothing about what it takes to establish a mechanism. Given the disambiguated picture argued for in Chapter 2, a standard EBM framework could be used to establish the existence of a mechanism by means of clinical studies. But to have any force as a programme for evaluating causality, EBM+ must offer a way to evaluate evidence obtained from mechanistic studies.

The problem of evaluation is a question about the feasibility of EBM+ as a methodology for causal evaluation. Williamson (2020, p.3) poses the question of whether "it [is] really practical to systematically evaluate mechanistic studies alongside clinical studies in medicine?" He identifies that an obvious way to meet this challenge is by providing "a good example of evidence assessment in medicine that appeals to EBM+ or something like it and that is clearly feasible" (ibid.). He finds a lack of examples of feasibility in the areas of intervention and disease assessment, but finds a positive example in exposure assessment. This positive example is found in the methodology of the International Agency for Research on Cancer (IARC), an organisation which evaluates evidence to determine whether certain compounds are carcinogenic. The methodology of IARC takes a similar approach to that of EBM+, as it explicitly evaluates evidence from mechanistic studies. Williamson argues that feasibility in exposure assessment translates to other areas of assessment, given that the EBM+ methodology is intended to be generalisable across domains within medicine.

A more direct way to answer the feasibility question is to use a novel framework built on EBM+ principles to carry out an evaluation of evidence from mechanistic studies. Some members of the EBM+ consortium have developed such a framework: '*Evaluating evidence of mechanisms in medicine*' (EEMM) (Parkkinen et al., 2018b). I introduce this framework in §3.2. This framework follows the EBM+ principles as in order to establish causation it requires establishing the existence of both a correlation and a mechanism. Moreover, when clinical studies do not suffice to establish a mechanism, EEMM requires evaluation of evidence from mechanistic studies. This framework shows that it is possible to construct an evaluative process from the EBM+ principles, but still does not offer evidence that EBM+ is feasible in practice. Additionally, in reviews of the framework there have been calls for testing of the evaluative guidance on a real-world example (Andersen and Kjaer, 2019; La Caze, 2019). The EEMM handbook provides some short worked examples, but these do not amount to a full evaluation that could be taken as evidence of feasibility.

The task of this chapter is to provide such a real-world example, adding to the evidence of feasibility of EBM+ that Williamson (2020) provides. To fulfil this task I will use EEMM, which I introduce in §3.2, to carry out a systematic review of the evidence for an intervention on Middle East respiratory syndrome (MERS). I introduce this case study in §3.3, and evaluate the evidence from mechanistic studies for a claim about the efficacy of the intervention in §3.4. Specifically, I evaluate the quality of evidence from mechanistic studies (§§3.4.1 and 3.4.2), and the status of a mechanism claim (§3.4.3). The review is used to demonstrate a basic kind of feasibility of EBM+: that one can carry out an extended review using a set of guidelines based on EBM+ principles. The review will then be used as a case study to evaluate EEMM in further chapters.

3.2 Introducing 'Evaluating evidence of mechanisms in medicine'

EEMM is a handbook that introduces the basic principles of the EBM+ approach, and provides guidance for evaluating evidence built on those principles. The guidance covers how to search for evidence from mechanistic studies, how to formulate mechanistic hypotheses and claims, and how to evaluate evidence of mechanism for the purpose of establishing causal claims. In this chapter I focus on the evaluation of evidence of mechanism. Causal claims are broken down into claims of efficacy (that the drug or exposure has an effect in a study population), and claims of effectiveness (that the drug or exposure has an effect in the target population). For ease of exposition I will focus on claims of efficacy in my review. The primary output of an evaluation is a *status* for a causal claim, where the status of a claim is a combination of the degree of confidence the evidence warrants in a claim, combined with the level of quality of that evidence. The levels and interpretation of the meaning of quality and status judgements can be found in Tables 3.1 and 3.2, respectively. Statuses of causal claims are arrived at by combining the statuses of correlation and mechanism claims, which are claims about the existence of a correlation and a mechanism, respectively.

A step-by-step procedure is to be followed to arrive at a status for a claim. Firstly, one should consider whether evidence from comparative clinical studies suffices to establish both a correlation and a mechanism claim (following the arguments presented in §2.2). This evidence base should be evaluated by an extant EBM framework (e.g., GRADE). If it is not of sufficient quality, then the evidence does not suffice to establish causation (recall from §2.2.3 that high-quality clinical studies can sometimes suffice alone to establish causation). If this evidence does however suffice to establish a correlation claim, then one should move on to evaluating the evidence for a mechanism claim, to which the next part of the procedure is dedicated. First one should evaluate the quality of mechanistic studies, and the quality of the total evidence for a mechanism claim. Then, the status of the mechanism claim is evaluated. Quality levels and statuses are what I call *evaluative judgements*. Judgements are arrived at by considering *evaluative criteria*. The criteria for evaluating quality can be found in Chapter 6, and the criteria for evaluating status can be found on page 83, of Parkkinen et al. (2018b). I will expand on the detail of these criteria in the relevant parts of the text of the review (§3.4). It suffices to say here that the criteria consist of a number of questions that are relevant to evaluating the quality of mechanistic studies, the quality of the total evidence for a mechanism claim, and the status of a mechanism claim. Finally, the minimum of the statuses of the mechanism and correlation claims then determines the status of the causal claim.

Quality Level	Interpretation
High	Further research is highly unlikely to have a significant impact on our confidence in the claim
Moderate	Further research is moderately unlikely to have a significant impact on our confidence in the claim
Low	Further research is moderately likely to have a significant impact on our confidence in the claim
Very Low	Further research is highly likely to have a significant impact on our confidence in the claim

Table 3.1: Interpretations of quality levels in EEMM. Adapted from Parkkinen et al. (2018b, p.26).

Status	Interpretation
Established	A claim is established when community standards are met for adding the claim to the body of evidence – i.e., for granting the claim and treating it as evidence for other claims In order to establish a claim, evidence must warrant a high level of confidence in the claim and this evidence must itself be high quality
Provisionally established	Moderate quality evidence warrants a high level of confidence in the claim
Arguable	The claim is neither established nor provisionally established, but evidence of at least moderate quality warrants significantly more confidence in the claim than its negation, or low quality evidence warrants a high level of confidence in the claim
Speculative	A claim is speculative if it falls into none of the other categories
Arguably False	The claim is neither ruled out or provisionally ruled out, but evidence of at least moderate quality warrants significantly more confidence in the negation of the claim than in the claim itself, or low quality evidence warrants a high level of confidence in the negation of the claim
Provisionally ruled out	Moderate quality evidence warrants a high level of confidence in the negation of the claim
Ruled out	A claim is ruled out when community standards are met for adding the negation of the claim to the body of evidence In order to rule out a claim, high quality evidence must warrant a high level of confidence in the negation of the claim

Table 3.2: Interpretations of statuses in EEMM. Adapted from Parkkinen et al. (2018b, p.27).

3.3 Case

To provide a defence of EEMM, and evidence of mechanism evaluation in general, I will carry out a systematic review of evidence of mechanism for a real-world intervention. I will first describe a disease and an intervention proposed for its treatment, and then evaluate the evidence of mechanism obtained from mechanistic studies relevant to this case.

Coronaviruses are enveloped RNA viruses from the *Coronaviridae* family. The family is named after a distinct morphological feature, namely, a fringe of spikes projecting from the surface of the virus that gives it the appearance of a solar corona (fig. 3.1). Four members of this family continuously circulate in human populations (HCoV-229E, HCoV-NL63, HCoV-OC43, and HKU1) (Milne-Price et al., 2014), though there are many more members that can infect other species (Masters and Perlman, 2013). In a process called *zoonotic transfer*, a mutation will sometimes allow a virus that only infects a certain species to become infectious to humans. Now termed a *novel virus*, it is often associated with an increase in virulence, since there will be no innate immunity in humans (Braciale et al., 2013). For example, although coronaviruses typically cause mild respiratory diseases in humans, from 2012 onwards a novel coronavirus has caused sporadic outbreaks of severe respiratory disease mostly concentrated in Saudi Arabia. Accordingly, the disease was called Middle-East Respiratory Syndrome (MERS) (Arabi et al., 2017). The virus is known as MERS-CoV.

The mechanism by which this disease is caused involves the virus binding to receptors on the surface of lung cells, then replicating inside the cells, leading to lung cell damage and severe respiratory syndrome (Arabi et al., 2017). This syndrome often leads to death, with a fatality rate of 35% observed as of 2017. There are currently no recommended treatments for MERS, but one strategy has been to repurpose drugs for other conditions. During the outbreak, a *combination therapy* of interferons (IFN) and ribavirin saw clinical use. This treatment was suggested on the basis of a combination of evidence from clinical and

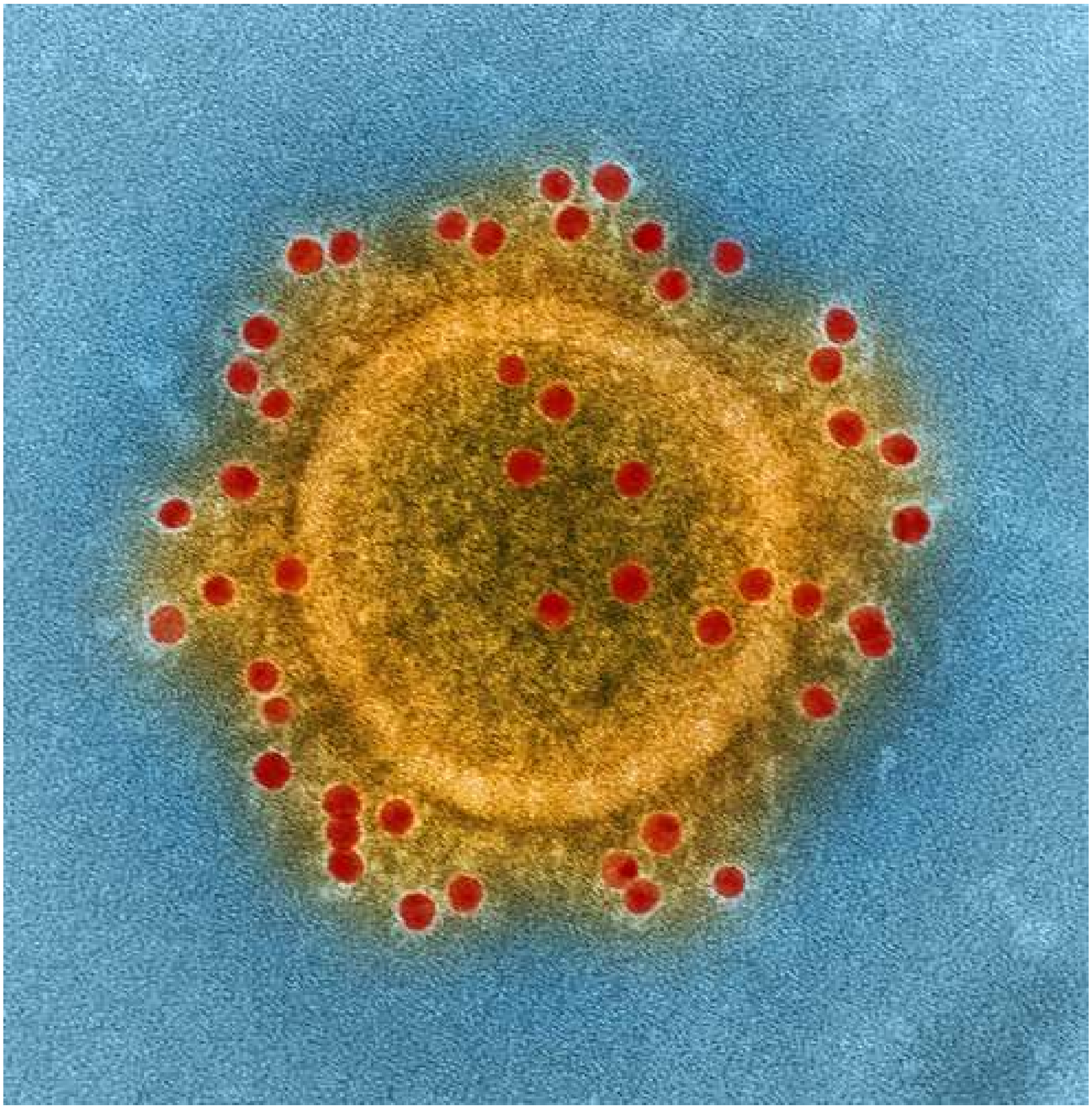


Figure 3.1: A colourised electron micrograph of MERS-CoV, the coronavirus that causes Middle East respiratory syndrome (MERS).

mechanistic studies. The proposed mechanism by which the treatment operates involves intervening on the viral replication component of the mechanism of disease. If combination therapy stops the virus from replicating, this should lead to prevention of further lung cell death and recovery from MERS.

There are a number of lines of evidence supporting the effectiveness of combination therapy. Firstly, ribavirin is a nucleoside analogue used as a broad-spectrum antiviral based upon its demonstrated efficacy *in vivo* and *in vitro* against viral replication for other coronaviruses (Coen and Richman, 2013). It was used against SARS, with uncertain effectiveness (Stockman et al., 2006), but it is typically used in novel viral outbreaks. Secondly, there is evidence of an underlying mechanism by which MERS-CoV replication is inhibited by interferons (IFNs) (Rabaan et al., 2017). IFNs are a family of cytokines that are a key component of the innate immune response against viruses. They induce an anti-viral state in infected and uninfected neighbouring cells (Feld and Hoofnagle, 2005). The problem is that viruses can produce accessory proteins to avoid or down-regulate this immune response by inhibiting the expression of IFNs (Yang et al., 2013; Menachery et al., 2017; Frieman et al., 2007; Devaraj et al., 2007; Iwasaki and Medzhitov, 2013). But this action may be countered by administering exogenous IFNs (Feld and Hoofnagle, 2005). Indeed, there is evidence that the replication of MERS-CoV and closely related coronaviruses is inhibited by exogenous IFNs (Hensley et al., 2004; Barnard et al., 2006; Wilde et al., 2013). Moreover, this inhibition prevents lung cell damage *in vivo* (Haagmans et al., 2004).

The major problem facing mono-therapies of IFN and ribavirin for treating MERS was that serum levels of either compound that could inhibit replication of the virus could not be obtained when administered to patients. However, there is also evidence of an underlying mechanism by which the combination of ribavirin and IFNs have a synergistic effect against MERS-CoV replication. When used in combination, the concentrations of ribavirin and IFNs needed to inhibit viral replication *in vitro* were lowered to clinically achievable levels (Morgenstern et al., 2005; Falzarano et al., 2013b). Moreover, a rhesus macaque model infected with MERS-CoV displayed signs of recovery when administered combination therapy (Falzarano et al., 2013a). Not only was survival prolonged in this model, but histopathological indicators obtained upon sacrifice of the animal showed reduction in cell damage relative to control animals. The studies that suggested synergistic anti-viral activity are those that I evaluate in the review below.

In accord with the EEMM process, we only move on to evaluating mechanistic studies once clinical studies are shown not to establish the causal claim. In this case, there is a paucity of clinical evidence supporting the use of combination therapy (Arabi et al., 2017). One problem is that the sporadic and severe nature of outbreaks means it is hard to recruit participants for trials. We are limited to case series and retrospective cohort studies, which are low-quality forms of clinical study. As a body of evidence these studies do not indicate a consistent benefit. However, there is a need to find treatments that may work. Evidence of mechanism may be instructive here. We cannot infer causation from the clinical studies alone, but finding that a mechanism operates will go some way towards making a causal relationship between combination therapy and recovery plausible. However, we cannot conclude that a mechanism exists merely by acknowledging that there is evidence from mechanistic studies of a mechanism linking combination therapy and recovery. We need instead to evaluate that evidence.

3.4 Review

In this section I present the details of a systematic review of the evidence from mechanistic studies for the effectiveness of combination therapy as a treatment for MERS. This evidence is used to evaluate a mechanism claim. In carrying out the review, I follow the stepwise guidance set out in EEMM and present a summary of the review in narrative form, which is standard procedure in systematic reviews. The final output of an EEMM review is supposed to be a status of a causal claim, which for this case would be of the form 'combination therapy causes recovery in MERS patients'. As the purpose of this review is to assess the feasibilty of the procedures for evaluating evidence from mechanistic studies, I will not provide a full evaluation of the correlation claim. This means I can only roughly indicate what the status of the causal claim might be, which I will do at the end of this section.

3.4.1 Quality of mechanistic studies

Each study is evaluated according to the guidance set out in Chapter 6 of EEMM. Briefly, the guidance states that we first evaluate the mechanistic studies from which we obtain the evidence of mechanism. Methods of these studies are assessed for how well they are understood, how similar the experimental system is to the target system, and how well the methods are implemented. This results in a level of quality for the evidence each study provides. I present the evaluation in a narrative form as this enables my reasoning to be made explicit. The evaluation was carried out by initially extracting data from the studies and identifying where the evidence did well or badly relative to the quality criteria. This raw evaluation can be found after the narrative review on pages 65 to 67 in Tables 3.3 to 3.5. I constructed these tables for the purpose of the review to make reasoning from the details of the evaluation to judgements on the strength of the evidence easier to carry out. Something like these tables would be a welcome addition to the handbook.

Falzarano et al. (2013a)

This study was an *in vitro*, cell culture study. It performed anti-viral assays on Vero and LLC-MK2 cell lines, using multiple techniques and measurements. The methods used are all well established virological methods, and the cell lines are fully characterised and standard in cell culture studies in virology (Condit, 2013). Issues with the study were *system dissimilarity*, *lack of control group*, and *no statistical analysis*. The system was dissimilar because MERS-CoV cell culture is far removed from the clinical course of MERS in humans. One mitigation of this issue is that the cells were derived from a rhesus macaque, which is an established animal model for MERS (de Wit et al., 2013; Yu et al., 2017). Lack of control group is a potential methodological problem for any study, as without one there is no way of knowing whether the effect observed is due to the intervention or due to chance. However, the study did implement some other kinds of controls by taking into account the resistance to ribavirin displayed by one kind of cells, namely, Vero cells. Using LLC-MK2 cells as well, which are not resistant to ribavirin, makes up for limitations in Vero cells by attempting to observe a consistent effect across the different cell lines. The absence of a statistical analysis of the results means that we do not know whether the reductions observed in viral assays were significant or not. This problem is likely the result of not using a control group. It is therefore uncertain whether the reductions were the result of the intervention or from the natural course of the infection. Finally, the study did not measure the effect of combination therapy with a quantitative method, whereas it did for monotherapies. The relative inaccuracy of the anti-viral assays used to measure the effect of combination therapy is a concern (Roldão et al., 2009).

The lack of controls and statistical analysis, and issues with measurement and system dissimilarity are partially mitigated by both the use of cell lines with known property differences, and the observation of a large synergistic effect. No other clear errors were committed. If a future study was to avoid committing these issues, it is likely that our confidence in the claim would change. Hence, this study is not high quality. However, the size of the effect makes it unlikely that we would observe such a large change in results that the synergistic effect would be entirely removed. We should therefore not expect a large change in our confidence in the claim, and for this reason, I evaluate the study as moderate quality.

Falzarano et al. (2013b)

This was a rhesus macaque animal model study. Six macaques were divided into two groups of three and infected with MERS-CoV. One group was administered combination therapy, and after 72 hours all macaques were sacrificed. Measurements were then performed through clinical examinations of dissected anatomy, virological and histopathological analysis of tissue, and viral load measurement by Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR). Quality issues with the study were system dissimilarity and inappropriate endpoints.

Drawing inferences from animal models to humans is more secure than from cell cultures, but is still subject to difficulties. Non-human primate models are considered the gold-standard of animal models for drug testing, as they are the closest in physiology to humans (Baseler et al., 2016). Rhesus macaques have previously been described as a sufficient model for MERS-CoV infection (de Wit et al., 2013; Yu et al., 2017). This means that important virological indicators and differences between humans and macaques are well characterised. However, this presents a problem for the study, as macaques only develop a mild form of MERS (Baseler et al., 2016; Yu et al., 2017). Hospitalisations that require combination therapy typically occur for the severe form of MERS. The extent of mild MERS is not known as it is a transient form of the disease that is unlikely to need treatment. Results showing effectiveness against this form of the disease may not strongly support the effectiveness of combination therapy in practice. Additionally, the endpoints measured in this study were inappropriate as they did not correspond to the clinical course in macaques. One major problem was that viral load was measured per tissue, but the results that were used to support the conclusion that viral replication was inhibited were presented as mean viral load across tissues. The variance of viral load reductions in treated animals was also high across different tissues. Viruses only affect some kinds of tissues, and in this study the tissue types that saw the most dramatic reductions in viral load are not those most effected in clinical manifestations of MERS. There is thus a disconnect between what was observed and the conclusions made on the basis of those observations.

System dissimilarity, a result of differences in presentation of MERS between macaques and humans, is partially mitigated by the use of a variety of accurate measurement techniques. However, common marmosets exhibit a form of MERS more similar to humans (Baseler et al., 2016). Testing of combination therapy in that animal model is likely to produce different results, thus changing our confidence in the claim. The quality of this study is therefore low-to-moderate.

Morgenstern et al. (2005)

This study was an *in vitro* cell culture study. It tested ribavirin as monotherapy on 6 cell lines (including 3 human cell lines), and combination therapy on 2 human cell lines (Caco2 and CL14), all of which were infected with SARS-CoV, the virus that caused the outbreak of *severe acute respiratory syndrome* (SARS) in 2003 (Stockman et al., 2006). Antiviral assay was performed, measuring cytopathogenic effect (CPE) by visual assessment, and then converting this into $TCID_{50}$, EC_{50}, EC_{75}, and EC_{90} values. All methods are well established in virology.

Quality issues with the study were system dissimilarity and lack of quantitative measurement. Two kinds of system dissimilarity are introduced in this study, over and above the usual problems with drawing inferences about effectiveness in humans from cell culture studies. One problem was the kind of cell line used. As noted above, different cell lines are effected to a variable extent by the same virus. This study used intestinal cell lines, whereas MERS primarily affects lung and to some extent kidney cells (Arabi et al., 2017). The other problem is that this study tested combination therapy on SARS-CoV infected cells. The two viruses are phylogenetically similar (see fig. 3.2), and treatments used in the SARS outbreak were used in the MERS outbreak. There are however known differences between the viruses, which pose potential problems for system similarity. In addition, CPE and $TCID_{50}$ are methods of measurement more prone to error (error rates of up to 35%) than quantitative RT-PCR (Roldão et al., 2009). This was not mitigated by the use of any quantitative form of measurement, or a more precise qualitative assay. This issue may be mitigated by the large degree of reduction in CPE observed.

The issues of system dissimilarity and absence of quantitative measurement are partially mitigated, as noted above. It is also the case that the nature of the cell lines is well characterised. However, without a control group, we cannot know whether the effect observed is real. The quality of evidence is low-to-moderate.

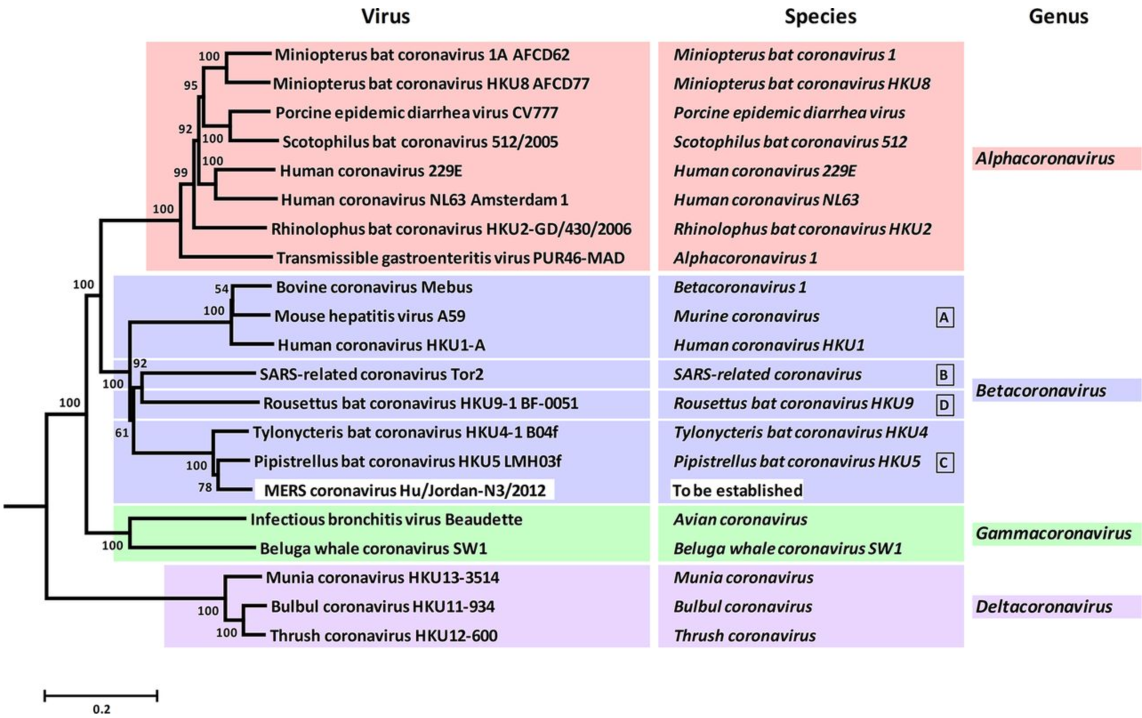

Virus
Species
Genus
Miniopterus bat coronavirus 1A AFCD62
Miniopterus bat coronavirus HKU8 AFCD77
Porcine epidemic diarrhea virus CV777
Scotophilus bat coronavirus 512/2005
Human coronavirus 229E
Human coronavirus NL63 Amsterdam 1
Rhinolophus bat coronavirus HKU2-GD/430/2006
Transmissible gastroenteritis virus PUR46-MAD
Bovine coronavirus Mebus
Mouse hepatitis virus A59
Human coronavirus HKU1-A
SARS-related coronavirus Tor2
Rousettus bat coronavirus HKU9-1 BF-0051
Tylonycteris bat coronavirus HKU4-1 B04f
Pipistrellus bat coronavirus HKU5 LMH03f
MERS coronavirus Hu/Jordan-N3/2012
Infectious bronchitis virus Beaudette
Beluga whale coronavirus SW1
Munia coronavirus HKU13-3514
Bulbul coronavirus HKU11-934
Thrush coronavirus HKU12-600
Miniopterus bat coronavirus 1
Miniopterus bat coronavirus HKU8
Porcine epidemic diarrhea virus
Scotophilus bat coronavirus 512
Human coronavirus 229E
Human coronavirus NL63
Rhinolophus bat coronavirus HKU2
Alphacoronavirus 1
Betacoronavirus 1
Murine coronavirus
Human coronavirus HKU1
SARS-related coronavirus
Rousettus bat coronavirus HKU9
Tylonycteris bat coronavirus HKU4
Pipistrellus bat coronavirus HKU5
To be established
Avian coronavirus
Beluga whale coronavirus SW1
Munia coronavirus
Bulbul coronavirus
Thrush coronavirus
A
B
D
C
Alphacoronavirus
Betacoronavirus
Gammacoronavirus
Deltacoronavirus
100
95
92
100
100
99
100
100
54
100
92
100
100
61
100
78
100
100
100
0.2

Figure 3.2: Coronavirus genera (de Groot et al., 2013). Note how closely related SARS-CoV and MERS-CoV are. Both are beta-coronaviruses and are closely related to other bat coronaviruses. On this phylogenetic tree SARS-CoV and MERS-CoV share a common ancestor only three splits back for MERS-CoV and two splits back for SARS-CoV.

Falzarano et al. (2013a)	Notes	Limitations
Methods	Antiviral assay using Vero (African green monkey) and LLC-MK2 (rhesus monkey kidney) cells infected with MERS-CoV. Cytopathogenic effect (CPE) measured by imaging cells. Viral load detection with RT-PCR (viral RNA levels). Viral protein levels assessed using western blot. Infectivity assay in vero cells.	
Understanding	Cell line types established for use in anti-viral assays. Detection techniques all established.	
System Similarity	Non-human cell lines. Vero and LLC-MK2 shown to be susceptible to MERS-CoV infection (Chan et al 2013). Also, Rhesus monkey is established as a model for MERS-CoV (de Wit 2013). However, kidney cells used here while target is lung cells.	
Surrogate Endpoints	Viral RNA and protein levels surrogates for viral inhibition. Established as surrogates in virology.	
Implementation		
Control	Not clear what the control is. No limitations referred to in paper. Vero cells resistant to ribavirin, so LLC-MK2 cells tested as well – similar effect noted. Vero cells unable to produce IFN – this allows effect of exogenous and not endogenous IFN to be measured. LLC-Mk2 cells do produce IFN – Combination therapy tested on LLC-Mk2 cells but no results displayed. Standard precautions for cell experimentation taken – constant temperature and PH etc.	/
Errors Committed	No explicit control group. No statistical analysis performed – no indication whether reductions were statistically significant or not. This is a problem as reduction for combination only given for TCID50 and viral nucleocapsid expression – unclear why quantitative viral load results not performed/recorded for combination treatment whereas they were for monotherapies. Reporting inconsistencies with no clear rationale – e.g. LLC-Mk2 cells tested but results not reported.	/
Results	IFN and Ribavirin alone reduced cytopathogenic effect, viral protein levels, and viral loads. Reductions also observed for combination therapy. Concentrations needed for same reductions as monotherapies reduced from 1000ug/ml to 62ug/ml for IFN and 200 to 25ug/ml for ribavirin	
Quality	Reductions in quality due to system dissimilarity, lack of control and no statistical analysis. No other clear errors committed. Lack of explicit control may be mitigated by use of two different cell lines with different and complementary properties. Synergistic effect large. Presence of controls make it moderate quality.	2

Table 3.3: Summary of evaluation of Falzarano et al. (2013a) by guidance in Parkkinen et al. (2018b, Ch.6). Quality criteria are grouped, with groups indicated by colors in the left column. Blacked out cells do not require evaluation, as they are either descriptive (Methods), or a category (Implementation). The 'Limitations' column is to be used as a heuristic to sum-up how significant any problems identified in the summary column are: each dash indicates one limitation; total limitations are found at intersection with the 'Quality' row. The total should be a guide for how low to rate the quality. This process is not found in EEMM and is borrowed from GRADE. It should not replace a full narrative review that shows the reasoning used to arrive at quality levels.

Falzarano et al. (2013b)	Notes	Limitations
Methods	Rhesus macaque animal model – 6 infected with MERS-CoV, 3 administered combination therapy and 3 untreated. Clinical examinations e.g. lung x-ray. Virological and Histopathological examination of necropsied animals. Histopathology – MERS-CoV antibody, tissues processed for immunohistochemistry using standard technology. Viral load measured by RT-PCR.	
Understanding	Macaque animal model previously described as sufficient for investigating MERS-CoV infection, including all important virological indicators e.g. virus shedding, replication rates and locations. However, macaques only develop mild disease similar to mild disease in humans – hospitalisations in humans only occur for severe form (Basiler et al 2016, Yu et al 2017). Techniques for histopathology and viral load established.	/
System Similarity	Non-human primate model – closest to human. Dissimilarities noted between human and model – see above. Main issue is the difference in disease profile.	/
Surrogate Endpoints	Genome quantification for viral load – appropriate. Viral load across all tissues taken as indicator of inhibition – may not be relevant as includes non-lung structures. Gross lung pathology for disease progressions – not clear whether this translates to survival. Both for survival.	/
Implementation		
Control	Untreated group stand as control group. No controls in place for difference in disease profile.	
Errors Committed	Animals euthanized at 72h – no long term follow up. Problem as De Wit 2013 proof of macaque model showed that viral loads were reduced without treatment in animals euthanized at 6 days compared to 3 days. Mean viral load across all tissues used as endpoint (see above) – MERS-CoV affects lower respiratory tract in macaques, and suspected upper respiratory tract in humans. Effect on bronchus tissue in macaques more relevant – data shows smaller reduction than mean reduction in this tissue. Largest reductions seen in trachea and non-lung tissues – not relevant to clinical course in macaque. May be mitigated by examination of lung pathology.	/
Results	Differences in gross lung pathology between treated and untreated animals. Significant differences in viral load – mean viral load across all tissues 0.81 log lower in treated animals. Some lung tissue showed no difference – see above errors.	
Quality	Reductions in quality for system dissimilarity and inappropriate endpoints. The main issue is the difference in severity of disease between humans and macaques. No stratification of reduction in viral load in relevant tissue also a major issue. The variety of measurements is a plus, as dramatic differences in lung pathology may mitigate some of the quality issues. Low to moderate quality – common marmosets are a more appropriate model for severe disease and testing in this model may produce different results.	4

Table 3.4: Summary of evaluation of Falzarano et al. (2013b) by guidance in Parkkinen et al. (2018b, Ch.6).

Morgenstern et al. (2005)	Notes	Limitations
Methods	Cell culture. Testing of Ribavirin alone on Vero, MA104 (both monkey), PK-15 (Pig), Caco2, CL14, HPEK (human). Combination tested on Caco2 and CL14. Anti viral assay – CPE by visual assessment, plus determination of TCID50 through visual assessment of CPE; inhibitory effects then determined for EC50, 75 and 90. Combination indices calculation to test for synergism.	
Understanding	All methods well established.	
System Similarity	Use of human cell lines a positive. Use of colon carcinoma cells introduces dissimilarity – MERS manifests as a respiratory disease, although intestinal and renal effects observed. MERS-CoV has been shown to infect intestinal epithelial cell lines (Chen et al.). Whether there are tissue differences in response to anti-virals is unclear – possible metabolic differences (see Fields). Combination therapy tested on SARS-CoV. Phylogenetically similar – known differences in multiple characteristics (inc. expected increase in sensitivity of MERS to IFN which has evidence for (Yen, Wilde).	/
Surrogate Endpoints	CPE for viral load. Reduction for viral inhibition. No other quantitative testing for viral load e.g. RT-PCR	/
Implementation		
Control	Control group is non-treated cells. Effect of moi explicitly controlled for by testing at different mois – no difference observed. No controls implemented for potential differences in cell lines – only one type of human cell line investigated for combination therapy sensitivity.	
Errors Committed	Not clear why only Caco2 and CL-14 cells tested on when multiple cell lines used for proof of infectivity. No statistical analyses performed to test whether reductions different to control. Due to dissimilarity of cell types, it is still plausible that viral inhibition would happen anyway.	/
Results	EC50, EC75 and EC90 all reduced for combination relative to monotherapy for either drug – ribavirin by factor of at least 10, IFN by 50-2000.	
Quality	Reduction of quality for system dissimilarity, lack of quantitative testing of viral load, and testing on only 2 cell lines. System dissimilarity is a minor issue – MERS-COV is more sensitive to IFN. Lack of rationale for testing only on 2 cell lines is a major issue. Human cell lines with known differences to human lung cells is a positive. Low to moderate quality.	3

Table 3.5: Summary of evaluation of Morgenstern et al. (2005) by guidance in Parkkinen et al. (2018b, Ch.6).

3.4.2 Quality of total evidence

Once the quality of the individual studies is assessed, the body of the evidence for the mechanism claim as a whole is then assessed for whether we have multiple studies that show consistent effects across similar and different kinds of methods. This would display a kind of robustness of results which may motivate judging the quality of the total evidence to be higher than if we judged on the quality of individual studies alone. When the same effect is robust to changes in background conditions it is more likely that the effect is real, rather than, e.g., a result of error. The quality of the total evidence of mechanism also requires combining the quality levels for each study. The weighting given to each study is determined relative to expert knowledge of the relevance of each study.

The combination of these studies do well on the 'stability of results' criteria. The same effect is observed across both different and similar kinds of methods. A number of different cell lines are used, including a variety of human cell lines. Finding similar results across cell lines increases our confidence that such results are not an artefact of the cell type used. A synergistic effect between the compounds that make up combination therapy was also observed in an animal model. So all types of study, using variable methods, show a decrease in concentrations of combination therapy needed to inhibit viral replication compared with mono-therapies of IFN or ribavirin. Carrying out a study in the future using a different kind of method is not likely to change our confidence in the claim because we already have evidence from different kinds of study demonstrating the same kind of effect. Moreover, across similar methods (the two cell culture studies), similar extents of synergy were observed.

The studies were individually found to be either low-to-moderate or moderate quality. Typically this was a result of dissimilarity of system. In only a few cases were reductions in quality a direct result of errors committed during the implementation of the study. High stability of results speaks in favour of the evidence, potentially raising its quality.

However, one major dissimilarity was that the only animal model exhibited a different disease course than in humans. This factor overrides the benefits accrued from observing stability of results. It is likely that results from different studies would be dissimilar if a more appropriate animal model is used in future studies. I therefore rate the evidence as low quality.

3.4.3 Status of mechanism claim

To assign a status to the claim that ‘there exists a mechanism linking combination therapy and recovery from MERS’, we are asked to consider what degree of confidence in the claim is warranted by the evidence, and combine that with the level of quality of the evidence. There are a number of questions that can raise or lower the status of a claim. These include, but are not limited to, considerations of the level of detail with which the mechanism is known, whether we have evidence for a crucial feature on which the mechanism depends, whether analogy is used to confirm the mechanism, or whether the mechanism is complex or unpredictable. Note that it is not clear whether particular criteria that can affect status do so by means of raising or lowering confidence in the claim or by affecting the total quality of the evidence. This point will be picked up on in Chapter 8, but here I will use a rough heuristic of altering the level of quality when the criterion is relevant to changes on future evidence bases, and altering confidence when the criterion is involved in indicating the truth of the claim.

In favour of holding a high degree of confidence in the mechanism claim is that we have multiple sources of evidence for the crucial feature ‘inhibition of viral replication’. Having evidence for ‘inhibition of viral replication’ should warrant holding high confidence in the claim. This is because if viral replication is inhibited then this stops the mechanism that ends in (death from) MERS. Principles of wider virology support this view (Condit, 2013). The evidence thus indicates that the claim is true. Reasons for lowering the status of the claim center on the use of analogy, which includes problems of system dissimilarity. The evidence for the claim comes from experiments on cell culture and an animal model with known differences to humans. We have to reason by analogy from confirmed mechanisms in virology to be confident that inhibition of viral replication in these systems will work in humans. A problem facing this use of analogy is the possibility of counteracting or masking mechanisms. This is another criterion in the process that results in lowering the status of the mechanism claim. Although we have evidence from cell culture and one animal model, it could still be the case that in humans there are other mechanisms that prevent the drug from working. Judging this to be so will result in a lowering of the level of quality of the evidence, as it will mean that future evidence could show that there are other mechanisms operating. I do not however think this is the case with combination therapy, and will illustrate why with a comparison to another potential MERS intervention, namely *mycophenolic acid*.

Evidence from mechanistic studies arguably supported the claim that there existed a mechanism by which mycophenolic acid inhibits viral replication (Milne-Price et al., 2014). Moreover, mycophenolic acid shows *in vitro* inhibition of viral replication for MERS at lower relative concentrations than did combination therapy (Hart et al., 2014; Chan et al., 2013; Cheng et al., 2015). There is also good evidence of a mechanism of action: Mycophenolic acid is a non-competitive reversible inhibitor of inosine monophosphate dehydrogenase, an enzyme involved in the synthesis of cellular purine nucleosides. This inhibition leads to a depletion of guanosine nucleosides, thus preventing viral replication, as guanosine is the precursor of guanine, one of the building blocks of RNA (Staatz and Tett, 2014). However, mycophenolic acid is typically used as an immunosuppressant drug in transplant patients, as it selectively depletes guanosine in T-cells (Staatz and Tett, 2014). So there was a possibility that it would initiate immunosuppressive mechanisms that would counteract its inhibiting effect on viral replication: “[I]ts immunosuppressive properties in vivo likely overshadow its direct antiviral effects, as no study has reported therapeutic benefit in animals” (Pierson and Diamond, 2013, p.787). Its effects on the immune system were thus likely to outweigh the beneficial effects on viral inhibition. A common marmoset model of MERS-CoV infection later found combination therapy to be ineffective (Chan et al., 2015). In sum, it seems likely that the putative mechanism linking mycophenolic acid with recovery in MERS patients was counteracted by the immunosuppresant mechanism. If this evidence were to be explicitly evaluated by EEMM, it would do badly on the 'are counteracting mechanisms likely' status criterion.

In contrast to this case, combination therapy was effective in an animal model. Even though the macaque model was disimilar in disease presentation, the fact that the drug did work provides some support to the claim that it is not counteracted. Moreover, combination therapy is a recommended treatment for Hepatitis C (for further details, see Auker-Howlett and Wilde (2020) and §2.2.4 of this thesis). This means its effects in humans are well characterised. In the mycophenolic acid case it is plausible that *in vitro* results would not extrapolate to use in humans, especially given our background knowledge. Here we have no reason to think a counteracting mechanism is plausible. Of course there is always the possibility that counteracting mechanisms do exist, but we have reasons to believe that they do not (I address this problem of *unknown complexity* in more detail in §4.3.2).

Another problem related to complexity is the ability of viruses to evade host immune responses. This ability was described in §3.4.1. It was identified early on that evasion proteins were present in the MERS-CoV genome (Wilde et al., 2013). Moreover, these evasion proteins were shown to inhibit the host immune response by downregulating expression of IFN. This evasion mechanism should not however pose a problem for predicting the effects of a treatment mechanism that involves administering exogenous IFN, which does not depend on gene expression to be effective. If the virus had the ability to evade IFN by blocking its downstream effects then this would pose a problem for the treatment mechanism, but this is not borne out by the evidence of the constitution of the viruses genome. Including genetics does add another layer of complexity, but what evidence we do have points to there being no effects on predictability by the complexity of the mechanism.

For these reasons, the use of analogy here is minor and not subject to major doubts about counteracting or masking mechanisms. Our confidence in the claim should therefore be moderate-to-high, as we have evidence for a crucial feature. There is also no reason to lower the quality any further than the level of quality assigned in §3.4.2. Combining the degree of confidence in the claim and the quality of evidence, I assign a status of *arguable* to the mechanism claim, on the basis of evidence from mechanistic studies. This is because there is low quality evidence that supports high(ish) confidence in the claim. A more conservative assessment might put it as *speculative*, but it is hard to see how it could be any lower.

3.4.4 Review conclusions

The point of this review was to demonstrate the practical feasibility of the EBM+ methodology for systematically evaluating evidence obtained from mechanistic studies. However, it is still worth giving an idea of what the status of the causal claim would be. After all, the point of the EBM+ programme is to evaluate evidence from clinical studies and mechanistic studies side by side. The evidence of correlation obtained from clinical studies supports a status for the correlation claim as *arguably false* or *provisionally ruled out*. Being more precise on the status of the correlation claim would require a full evaluation of the evidence for it, which would likely require the use of an existing evaluative framework, e.g., GRADE. But a coarse evaluation can be made on the basis that according to most EBM frameworks, the methodologies used in the clinical studies typically produce low to very-low quality evidence. Moreover, while the best quality studies suggest that no correlation exists, those same studies are beset by numerous quality issues. It is therefore likely that our confidence would change in the light of new evidence. For example, an analysis of study participants shows that, on average, they had a high-degree of co-morbidities. This could explain the high degree of mortality rather than just the effects of MERS. Additionally, in one study that found no correlation, a statistically significant decrease in mortality was found at 14 days, but not at 28 days. However, the p-value at 28 days was 0.054, which is only 0.004 over the typical significance threshold of 0.05. This alone should cast doubt on concluding that no correlation exists, and this doubt is compounded by the numerous methodological design and implementation problems.

In sum, one should hold low confidence in the correlation claim, warranted by low quality evidence. This would mean a less than *speculative*, but higher than *ruled out*, status. Without a full evaluation, only a coarse evaluation of an *arguably false* or *provisionally ruled out* status can be given to the correlation claim. As the minimum of the statuses of the correlation and mechanism claims gives the status of the causal claim, it is *arguably false* or *provisionally ruled out* that combination therapy causes recovery in MERS.

3.5 Conclusion

In this chapter I used the EEMM framework to evaluate the evidence from mechanistic studies for the claim that there exists a mechanism linking combination therapy and recovery from MERS. I argued that this claim was made *arguable*, as the low-quality evidence warranted holding high confidence in the claim. This review responds to one worry about the feasibility of EBM+ as a methodology for evidence evaluation. For EBM+ to be a more complete epistemology of causal evaluation, it must be possible to construct an evidence evaluation framework on the basis of its principles. In particular, the principle that evidence from mechanistic studies can be explicitly evaluated and combined with evidence from clinical studies. I demonstrated this by carrying out a review using the EEMM framework. This review was able to judge a status of a mechanism existence claim, and indicated how this would be combined with a full evaluation of clinical studies. This chapter also introduces a case study by which to evaluate EEMM in greater depth in the following chapters.

Chapter 4

Practical problems

4.1 Introduction

In the previous chapter, I demonstrated the feasibility of EBM+ by carrying out a systematic review of evidence obtained from mechanistic studies. The review supports the notion that one can perform an evaluation. I now turn to defending the core part of the evaluative process. A novel evaluative framework is likely to not be the finished article. To improve the framework one can analyse it on a conceptual (Chapters 5 and 7), and practical level. A full defence of the EBM+ approach to evidence evaluation must address both levels.

In this chapter I address some specific and general practical worries about i) evaluating evidence from mechanistic studies, and ii) establishing a mechanism on the basis of that evidence. I first identify a worry that arises from the application of EEMM to the MERS case study (§4.2). This worry is specific to the process and concerns what happens when a correlation is not established. The process says we should not proceed to evaluate evidence from mechanistic studies. I argue that this part of the process should change. I then answer some general, practical worries about evaluating evidence of mechanism. One worry is *method heterogeneity* (§4.3.1). This is the worry that because methodologies of mechanistic studies are more diverse than that of clinical studies, conducting a detailed systematic review of evidence obtained by means of mechanistic studies may be too difficult. I argue in response that my review, conducted in accord with the guidance of EEMM, is of comparable detail to reviews conducted on clinical studies. Another worry is *complexity and incompleteness* (§4.3.2). This is the worry that as most mechanisms are complex, this fact may preclude holding sufficient knowledge of the detail of a mechanism to establish it. I again appeal to my systematic review to show how parts of the process head off this problem. I identify that responses to these problems rely on expert judgement, which may invoke concerns about the influence on judgements of subjectivity. I delay responses to this worry to Chapters 6 to 8, as they are worries about the *malleability* of the methods of EBM+, which is a separate and significant charge that EBM+ and EEMM must be defended against.

4.2 What to do when there is no correlation to explain

Consideration of the conclusions of my review raises a potential problem. The status of the causal claim was less than speculative because the status of the correlation claim was less than speculative. One might think this makes the case study invalid, as evidential pluralism requires a correlation for the mechanism to explain. The process, taking its motivation from the evidential pluralism thesis, only instructs evaluators to consider mechanistic studies when a correlation claim is at least arguable. Why then should I have proceeded to evaluate the mechanism claim, when there is likely no correlation to explain?

One reason to pursue the evaluation invokes Illari's distinction (§1.2). When we distinguish by the objects of evidence, and not by the methods by which evidence is obtained, we see that clinical studies can provide evidence of mechanism, and mechanistic studies can provide evidence of correlation. EEMM provides some instruction on how evidence from clinical, respectively mechanistic, studies can boost mechanism, respectively correlation, claims (Parkkinen et al., 2018b, pp.92-94). For example, clinical studies may not suffice to establish a correlation because all plausible confounders were not controlled for. If high-quality evidence from mechanistic studies rules out these confounders, then it can help to boost the status of the correlation claim to established. Thus far the status of the correlation claim in the MERS case has been judged solely on the basis of clinical studies. It may be possible to make a case that evidence from mechanistic studies can 'boost' this claim by providing evidence of correlation.

The clinical studies that provided evidence of correlation are beset by numerous problems beyond the general problems faced by their methodological design. There are many potential inhibitors, where inhibitors are factors that, instead of making a correlation spurious as a confounder would, prevent a real correlation from being observed. Inhibition can be thought of as the masking problem applied to correlations. One example of an inhibitor present in these studies is inadequate time from infection to treatment. Evidence from mechanistic studies can explain how a factor is an inhibitor. Regarding time-to-treatment, in SARS, delayed IFN expression and late IFN administration was associated with stronger viral infection and worse outcomes in two mouse models (Channappanavar et al., 2016; Haagmans et al., 2004). Corman et al. (2015) found that MERS viral dynamics closely match SARS. It is also the case that clinical manifestations of MERS seem to exceed SARS in their rapidity (Zumla et al., 2015). This is supported by pharmacodynamic measurements that showed patients with more severe disease had higher levels of virus in the blood, a sign of severe infection. So, evidence from mechanistic studies show that treating MERS early is vital. Otherwise patients progress to a more severe form of the disease. Unfortunately, in the clinical studies, the time to treatment was often greater than the median time from onset of symptoms to death. For example, in Al-Tawfiq et al. (2014), the median time to treatment was 18 days, while the median time to death in all patients is 11.5 days (Zumla et al., 2015). If those studies did administer the intervention earlier, then it is plausible that we would have observed greater rates of recovery. So a case can be made that the status of the correlation claim would be greater on the basis of both mechanistic and clinical studies than on clinical studies alone.

There are two problems with this analysis. One is that the mechanistic studies cited here are not those included in the evaluation. Including them would be legitimate according to the process, as it is the total evidence for a claim that counts, not just a particular subset of the evidence. But EBM+ would require this evidence to also be explicitly evaluated, which is not carried out here. Now such an evaluation is practically feasible, as demonstrated by my review. To simplify things, I claim that it is plausible that these mechanistic studies could boost the correlation claim, making the review worth carrying out. However, the second problem is that the guidance in EEMM on 'boosting' claims only considers examples where mechanistic studies boost correlation claims to a status greater than speculative. It is still the case that EEMM would only consider evaluating mechanistic studies when a correlation claim is (partially) established. It is not clear that the correlation claim in the MERS case would be boosted to greater than speculative. Therefore, it is still doubtful that a review is useful in this case, according to the EEMM process.

I claim however that this reflects badly on EEMM, rather than on this case study, as the problem identifies an inadequacy in the methodology. Even when the status of a correlation claim is considered to be not above speculative, there is still a need to evaluate evidence obtained by means of mechanistic studies. This is because evaluation of evidence is important for both decisions on what interventions to recommend, and on the *direction of future research.*

To see why, first we must acknowledge the reasons for why a claim is given a status less than speculative but greater than ruled out. A status judgement involves assigning a degree of confidence to a claim *and* a level of stability for that confidence. Confidence is stable, respectively unstable, when it is unlikely, respectively likely, that future evidence will change the degree of confidence in the claim. In the MERS case, the clinical studies indicate that there is no correlation, so arguably one's confidence in the negation of a correlation claim should be high. This makes the status less than speculative. But the claim is not ruled out, as the quality of the evidence that warrants this degree of confidence is low. This means that the high degree of confidence in the negation of the correlation claim should be unstable. Future evidence bases are possible that, for example, do not use low-quality study designs, do control for inhibitors, or do find highly statistically significant results. Moreover, improvements in these areas would significantly change our confidence in the claim.

The next thing to acknowledge is that details of evidence evaluations are important for decision making. Moreover, they are important for different kinds of decisions. Typically, decisions are made about what interventions to recommend. As argued in §2.3.2, one need not establish a claim to commit to an action such as recommending a treatment. But it is plausible that decisions on recommendation would at least require a causal claim to be above *speculative*. However, decisions also need to be made about the direction of future research. This is particularly pertinent in the MERS case where there are currently no recommended treatments. And the reasons for which these decisions are made will be different to those given for decisions about what treatments to recommend. Arguably, decisions on which way to direct future research will consider whether evidence bases are currently conclusively, or inconclusively, in favour or against a causal claim. If an evidence base for the efficacy of an intervention is inconclusive, then it is more likely that further research on that intervention needs to be carried out than if the evidence base was conclusively for or against its efficacy.

In the MERS case, the correlation claim is not ruled out, and this is because of a judgement that it is likely that a future evidence base would result in a significant change in confidence in the claim. This sort of evaluation is more informative than one that observes 'no correlation' and concludes 'no causation', which is the point of detailed, systematic reviews. However, resources are limited and pursuing every claim that is not ruled out would be costly. Efficiency considerations motivate what I propose here, which is that we should evaluate mechanistic studies when a correlation claim is less than speculative. Finding that a mechanism claim is more than speculative may indicate that future research is worth pursuing. Equally, if *both* the correlation and mechanism claim are less than speculative then research is likely not worth pursuing. However, as long as neither of the claims are ruled out, further research is to some extent plausible to pursue. It is just that research is more likely worth pursuing if at least one of the claims is more than speculative. My approach thus provides a way to identify which avenues of research are worth pursuing, given that we have limited resources. Therefore, EEMM should include guidance on evaluating evidence of mechanism from mechanistic studies when a correlation claim is *neither (partially) established nor ruled out.*

One might object that EEMM is not in the business of directing future research. It is certainly the case that there is nothing systematic on decision making about treatments in EEMM, let alone about future research. Moreover, evidence of mechanism is already part of directing future research: as Solomon (2015) argues, evidence from mechanistic studies is more at home in the discovery process, where it can suggest treatments for the methods of EBM to test. The point is that this addition to EBM+ is already part of EBM. And what research is pursued within an EBM methodology depends already on the statuses of causal claims. However, central to EBM+ is the contention that causal claims are better evaluated when evidence of mechanism from all sources is explicitly evaluated. This includes judgements on quality and status. And it is not clear that future research decisions under an EBM framework, where they do consider mechanisms, do so in a way that explicitly and systematically evaluates evidence from mechanistic studies. Therefore, an addition to the process that is motivated by concerns about decisions on future research improves on the methodology of EBM in the same way that EBM+ proposes to with respect to evaluating causality. EBM+ is, after all, capturing all aspects of EBM and adding to them the explicit evaluation of evidence of mechanisms alongside and on a par with evidence of correlation. This should also include the part of EBM concerned with making decisions about future research.

Whether it is feasible to add to EEMM in this way is another matter. A potential difficulty one might have involves what the outputs of such a process might be. The final claim would be 'should future research be carried out'. I have argued that the answer to this question is *yes* when the status of a correlation claim and a mechanism claim are both higher than *ruled out*, and *no* otherwise. Additionally, when at least one of the claims is higher than speculative, then future research into this area should be prioritised over other areas where both claims are below speculative. One might wonder then whether it would be useful to apply statuses to this claim such that it can be *established*, *arguable*, *provisionally ruled out*, and so on. This would certainly make such an evaluation much more informative than a yes/no answer, which is the aim of prioritising certain avenues. However, this more fine grained status output would require equally fine grained interpretations, instead of just 'when the correlation claim and mechanism are both less than speculative and higher than ruled out'. And borrowing from EEMM and saying that the status of the future research claim is the minimum of the correlation and mechanism claims would miss the point of how the claims interact relative to this process. For example, the fact that the status of a correlation claim is *arguable* and the mechanism claim *provisionally ruled out* would mean that future research is also *provisionally ruled out.* But on my proposal it should actually be quite plausible that research should continue, as the correlation claim indicates that it is worth pursuing research, especially in light of the fact that one's confidence in the mechanism claim is unstable. So the application of EEMM's approach to combining statuses to evaluate a causal claim would be inconsistent with the reasons for pursuing this 'future research' proposal.

A resolution to this question would require more work on what it takes to motivate future research. The way in which correlation claims and mechanism claims interact to establish causality is motivated by evidential pluralism. By contrast, there is no philosophical theory supporting this 'future research' idea. It is likely however that there is some coherence to the reasoning used when directing future research in practice. Analysis of this area of medicine would aid in resolving this difficulty. It suffices here to note that this approach has motivation, if indeed it does not have a full development at present.

4.3 General worries about evaluating evidence from mechanistic studies

In Chapter 2, I proposed that a proponent of EBM may argue that evidence from high-quality clinical studies can establish the existence of a mechanism, but that evidence from mechanistic studies cannot. EBM+ makes no distinction between the kinds of methods used to obtain evidence of mechanism, so such a view would remove the point of the EBM+ programme. EBM evidence evaluation frameworks would thus remain in their current form. To support this view on mechanistic studies, a proponent of EBM could cite issues that are specific to establishing a mechanism on the basis of mechanistic studies. I argued such issues were no problem for the epistemological thesis that motivates EBM+, evidential pluralism, as an analysis of the epistemology of causation in medicine. I concluded that establishing a mechanism by any means should be *in principle* possible, but *in practice*, establishing a mechanism on the basis of mechanistic studies may be quite difficult. This is a concern about the evaluation of evidence from mechanistic studies, and in this section I address two general worries about it.

4.3.1 Method heterogeneity

The first issue arises because the methods utilised in mechanistic studies are more varied than those used in clinical studies. The worry is that a high degree of *method heterogeneity* could make evaluation of evidence obtained by these methods difficult. Each kind of method will have its own particular design features, where the design is such that the study can control for potential errors. Evaluating whether this has been successful requires guidance on each kind of control. In contrast, systematic guidance for clinical studies has a more restricted set of methods, and so a much more restricted set of design features to assess. Moreover, assessing evidence from clinical studies is still no easy task, implying it will be much harder for mechanistic studies.

One can see evidence of the heterogeneity problem in my review. The three studies used two broadly different methods, namely cell-culture methods and an animal model. Moreover, there was also within-method diversity. The two cell culture studies used different kinds of cells and different kinds of measurement techniques. For example, Morgenstern et al. (2005) utilised a number of cell lines derived from different animals: the monkey derived Vero and MA104 lines; the pig derived PK-15; and, the human derived Caco2, CL14, and HPEK. The animal model used only one kind of non-human primate, a rhesus macaque. But, as pointed out in the review, common marmosets would be a more appropriate model to use. So even within the broad method of animal models there are methodological intricacies likely not found *within* particular methods in the clinical sciences. Each technical difference across methods introduces new errors that must be assessed for. The problem facing EEMM is providing evaluative criteria that can generalise across all these kinds of methods, while also allowing evaluators to carry out a thorough assessment of particular kinds of method. Contrast this with a criterion instructing one to evaluate whether a clinical study has implemented randomisation, a particular design feature that is relevant to all clinical studies. It seems unlikely such a precise criterion can be fashioned for mechanistic studies. The worry is that an evaluation of mechanistic studies will be too general to make a justified judgement on the status of a mechanism claim.

To meet this challenge, EEMM must be able to provide guidance instructive enough to apply to all methods, but that still results in an evaluation that has a level of detail comparable to that found in evaluations of clinical studies. Degree of detail is a somewhat vague notion, but as a comparison, consider a systematic review into the treatments for SARS by Stockman et al. (2006). The supplementary information for this review contains evaluations of clinical studies and mechanistic studies. Importantly, the evaluations of each kind of study are significantly different on account of their extent of detail. Where clinical studies are concerned, a typical summary of an evaluation is:

> "[The] control group was patients admitted at the same time and treated with corticosteroids only or those who declined interferon treatment. Lack of randomization and small sample size makes effect of treatment difficult to conclude." (Stockman et al., 2006, Table S7)

Contrast this with an equivalent summary of evidence from mechanistic studies:

> "Pegylated interferon-a significantly reduces viral replication, observed a dose dependent effect." (Stockman et al., 2006, Table S5)

In the first quote, study design and implementation issues are detailed. In the second quote, only the results of the study are reported. This pattern is seen throughout the SARS review. The difference of detail can be summed up in these two schemas: the *detailed schema* is along the lines of 'X detail means that inference Y is/is not justified'; the *support schema* is along the lines of 'Evidence E supports conclusion C'. For my review to be considered detailed it must be in the form of the detailed schema. This is because it can be used to justify inferences about the overall strength of the evidence, rather than to just communicate what the conclusions are.

The review carried out in Chapter 3 does provide an extent of detail comparable to that achieved in reviews of clinical studies. Moreover, this review was performed on a variety of method types. The degree of detail extracted using only the evaluative criteria contained in EEMM can be seen in the narrative summary in §3.4.1 and the raw evaluation in Tables 3.3 to 3.5 on pages 65 to 67. Moreover, phrasing was of the form of the detailed schema: "X detail means that inference Y is/is not justified". For example, a judgement of low-quality was given to the study that tested combination therapy in a rhesus macaque model (Falzarano et al., 2013b), as:

> Common marmosets exhibit a form of MERS more similar to humans (Baseler et al., 2016), and testing of combination therapy in that animal model is likely to produce different results, changing our confidence in the claim.

Extraction of this detail was guided by the evaluative criteria that instruct evaluators to assess whether the experimental system used is similar to the target system. Such a criterion is generalisable across domains, but allows a detailed review of a specific methodology.

One might doubt that the review was detailed *enough*. In EEMM there is significant latitude given to reviewers to use their domain expertise to apply the criteria to specific studies. Given the complexity of some methods, it is plausible that some factors that matter for evaluating the quality of a study may be missed. Instructing reviewers, for example, to assess 'whether typical errors have been committed', which is one of the quality criteria, may be too vague. Whether all errors are identified will depend on the knowledge and biases of the evaluator. This problem may be ineliminable due to the nature of evidence evaluation. Indeed, expert judgement is likely to play a part in any evaluation. To mitigate this, EEMM does provide structure to guide expert reasoning, and this may be all that is achievable. Moreover, the influence of expert judgement on evidence evaluation will be a repeating theme through this thesis. This is a concern about the *malleability* of evidence evaluations, rather than just feasibility, and I will address specific concerns with it in Chapters 5 and 7. For now, it is enough to say that method heterogeneity as such does not pose a problem for establishing a mechanism on the basis of mechanistic studies, even if worries about expert judgement may end up casting doubt on the veracity of the claim.

Compounding this issue with expert judgement may be a wider problem with the biomedical sciences, which results from it being a less cohesive domain than clinical science. Methodological heterogeneity means that the kind of textbook approach to the methods used in clinical science is not found in the biomedical sciences. There is a wealth of methodological information relevant to mechanistic studies in publication, but it is not centralised, systematic, or accessible in the way guidance on the methods of clinical science is. For example, one can consult textbooks (Fletcher et al., 2012) or evaluative frameworks (Guyatt et al., 2011b) and find most if not all of the information you would need to evaluate evidence obtained from clinical studies. General guidance on the kinds of assay used in virology can be found (see, e.g., Condit (2013)), but specific information about, e.g., error rates, may be published piecemeal in journals (see e.g., Roldão et al. (2009)), if it is at all.

A resolution of this problem would be external to EEMM, but something that could positively add to the EBM+ programme. This would be to produce standardised guidance on methodologies used throughout the biomedical sciences. If the claims and methodology of EBM+ are defensible, and my contention in this thesis is that they are, then there is motivation for providing the kind of centralised, systematic, and accessible methodological guidance that is a mainstay of standard EBM. This would be a natural extension to the EBM+ project.

4.3.2 Complexity and incompleteness

Mechanisms are complex due to the high number of components and relations of which they are composed, and an organisation of these components and relations that is often not linear. This complexity can often lead to difficulties in prediction, and to *epistemic incompleteness*. It is important to distinguish between the metaphysical claim of complexity, and the epistemic consequences of that complexity. Mechanisms themselves are not incomplete, but our evidence may be if we so happen to have none for some component of a mechanism. This incompleteness of evidence can lead to unpredictability, but complexity itself can lead to unpredictability as well. Such arguments were addressed as rejections of evidential pluralism in Chapter 2, but they remain issues for establishing the existence of a mechanism on the basis of mechanistic studies. Complexity of a mechanism means that i) in order to establish a mechanism the amount of evidence needed is likely to be very large, and ii) it is difficult to be sure that at any one time we have all the evidence needed to establish the mechanism.

A simple move to counter this issue would be to again acknowledge the distinction between the details and existence of a mechanism. While a mechanism may be highly complex, we need not require all of the details of this mechanism to establish its existence. This move runs into immediate problems once we are considering the practical process of establishing a mechanism. Establishing the existence of a mechanism on the basis of mechanistic studies will depend on what we consider sufficient detail, as individual components will need to themselves be established. We may not require all the details, but we will require some of the details, and we might be concerned that knowing when we have the requisite detail is too difficult.

EEMM tries to counter this problem with specific criteria, inclusion of which is intended to put limits on what counts as acceptable detail. For instance, evaluators are asked to consider whether the mechanism exhibits such complexity that its overall behaviour is very unpredictable. Just having knowledge that the mechanism is likely to be highly complex will cause evaluators to lower their confidence in the claim. But having evidence that this complexity is of little concern would mitigate any reduction in confidence. In the case of MERS and combination therapy, the treatment mechanism is not highly complex. Consideration of potential masking mechanisms heads off concerns that there are other mechanisms adding to the known complexity. With respect to incompleteness, evaluators are asked to consider a number of questions: Is the mechanism known in some detail?; Are most of the crucial features of the mechanism known and understood?; Is it plausible that the behaviour of the mechanism crucially depends on just some components or organisational features? In my systematic review, these criteria were explicitly considered. Inhibition of viral replication is a crucial feature in a mechanism to treat MERS. Details such as 'how the virus enters into a host cell' are not as important for this mechanism, as the treatment combats MERS-CoV replication through other means. Other experimental MERS treatments do target viral entry sites, such as the use of monoclonal antibodies to block the DPP4 receptor, which MERS-CoV uses to gain access to host cells (Rabaan et al., 2017). Evidence for this component would be required if the mechanism concerned the action of monoclonal antibodies against MERS. Given that we know the IFN mechanism in some detail, the evidence required in the case of combination therapy is that IFN inhibits viral replication, and does so synergistically with ribavirin.

A worry about this response is that it may not deal with *unknown* complexity and incompleteness. One could object that using our current evidence to tell us whether that same evidence is complete is not a legitimate strategy. What we need instead is independent reason to think that our evidence is complete enough. However, using the notion of *crucial features* is one way of legitimately implementing the strategy of using our current evidence to identify incompleteness. Whether there are missing components does not matter if we have evidence for a component on which the operation of the mechanism depends. But this strategy invites a regress. How do we know we have all the crucial features? It is plausible that many mechanisms have more than one crucial feature, and identifying what counts as a crucial feature will depend on evidence. So a problem of *unknown crucial features* arises when one utilises current evidence of crucial features to mitigate the incompleteness problem.

Instead of appealing to our current evidence, one might appeal to background knowledge to mitigate this objection. Background knowledge provides grounds for what sort of mechanisms, and what sort of features of mechanisms, can be expected in a particular context. This background knowledge can instruct on whether there are likely to be more crucial features. In the MERS case, more crucial features are not likely, given the importance of inhibition of viral replication in anti-viral treatments. For an intervention on cardiovascular disease, given the complexity of cardiovascular mechanisms, it is likely that there will be many crucial features. Evaluations of evidence for such an intervention will require much more detail than is required in my case study.

This sort of strategy is employed in the discovery of mechanisms. Craver and Darden (2013) argue that researchers use constraint based reasoning to gain complete knowledge of mechanisms, where those constraints are imposed by current background knowledge. For example, in the discovery of the mechanism of the action potential, physical theory was used to limit what kind of components were possible given current experimental results. This reasoning process identified ion channels as the most likely candidate to be the component central to propagating the action potential (Craver, 2006). Later on in the discovery process, when the mechanism was known in some detail, thermodynamic considerations pointed to the possibility of a missing feature. The mechanism could not account for how the ion channel overcame energy constraints that should have prevented it from opening (Catterall, 2010). A feature that resolved this problem was sought for and subsequently discovered (Chen et al., 2010). The example shows how there may always be detail to find and how constraint based reasoning guides research. But it also shows how we can identify when the detail we seek matters for whether the mechanism operates or not. Plausibly, it does not matter how the ion channel opens, if what we want to know is whether there is a mechanism by which an ion channel opens to allow the action potential to propagate. And this was not in doubt. I suggest that this *constraints based reasoning* approach can be applied to working out whether there are unknown crucial features.

Background knowledge can also be used to establish whether unknown complexity poses a problem to establishing a mechanism. Typically, this process will involve using knowledge of the complexity of other mechanisms in some domain to inform judgements about the mechanism under evaluation. In virology, there is a wealth of mechanistic knowledge about how viruses operate. Virology has been an established domain for a significant amount of time, and has arguably seen some of medicine's greatest successes (e.g., eradication of Smallpox, near-eradication of polio, end of the AIDS epidemic in developed nations). The mechanisms will of course not be simple: in virology the amount of relevant components and relations will be high, and not all mechanisms will operate linearly. But, previous experience will have shown the sort of negative effects we should expect complexity to have on predictability. For example, the ability of viruses to evade host immune responses is well known, and this same phenomenon has posed problems for otherwise well designed treatments in the past. Researchers can mitigate this problem through the rapid sequencing of viral genomes, allowing identification of the potential existence of proteins involved in 'viral evasion'. In the MERS case, it was identified early on that evasion proteins were present in the MERS-CoV genome (Wilde et al., 2013). The effects of such proteins can be investigated to inform judgements on whether there would be an effect on the complexity and predictability of any treatment mechanism. My review judged that this would not be a significant effect, as the viral evasion proteins worked on the expression of *endogenous* IFN, and the treatment involved administering *exogenous* IFN.

One could object that appealing to other mechanisms, such as in wider virology or genetics just increases the degree of complexity, making it harder to predict effects. In response, it might be the case that ruling out entirely the negative effects of incompleteness and complexity may be unachievable. But, we can rule out enough in some cases to hold a reasonably high degree of confidence in a mechanism claim. There will always be some degree of doubt in any scientific claim. In a similar vein, incompleteness of evidence is always a worry in any discipline. But we can still make informed judgements about whether it may be enough of a worry to make us consider the claim implausible. Establishing is a fallible status, so claims can always be overturned. But this should not count as a reason for why a mechanism can never be established.

4.4 Conclusion

In this chapter I have responded to a number of concerns about the practice of evidence evaluation on EEMM. One concern arose through conducting the systematic review: what to do when a correlation is not established. I argued that it is worth carrying out a review when evidence from mechanistic studies does not suffice to make a mechanism claim at least *arguable*, if it is plausible that evidence from both mechanistic and clinical studies supports a higher status for the causal claim than on clinical studies alone. Moreover, even when it is not the case that mechanistic studies boost the correlation claim above a speculative status, I argued that it is still worth evaluating evidence from mechanistic studies when the correlation claim is not ruled out. This is because an evaluation of evidence from both kinds of methodology is better for directing future research than on only one kind. I identified that more work would need to be done to flesh out this proposal. Two concerns derived from a worry that carried over from the defence of evidential pluralism in Chapter 2, namely, that it is not possible to establish a mechanism *in practice* through evidence obtained from mechanistic studies. I responded to the *method heterogeneity* worry by arguing that my review of evidence from a variety of different mechanistic studies was conducted to an extent of detail comparable to systematic reviews carried out on clinical studies. I then argued that *complexity and incompleteness* worries can be mitigated by the evaluative criteria found in EEMM. However, these criteria should be supplemented by more guidance on reasoning strategies to use when one is confronted with the problems of *unknown complexity and incompleteness.*

Chapter 5

Conceptual problems

5.1 Introduction

An evaluation by EEMM results in one main outcome: a judgement on the *status* of a causal claim, arrived at by making judgements on the statuses of both correlation and mechanism claims. To make a status judgement one must also make a judgement on the *quality* of the evidence used to support those claims. This is because of how one should interpret the meaning of each kind of status. Each kind of status is interpreted as a level of quality of evidence that warrants holding an extent of confidence in a claim. For example, when one judges the status of a claim to be *provisionally established*, this means that 'moderate quality evidence warrants a high level of confidence in the claim'. So, a status judgement requires first evaluating the quality of evidence. Further, each level of quality is interpreted as the extent to which one's degree of confidence in the claim is stable in the light of future evidence. Ultimately then, evaluating the evidence leads to a judgement based on the extent and stability of one's confidence in a claim. Together, I call judgements on status and quality, *evaluative judgements*. Interpretations for all quality and status judgements can be found in Tables 3.1 and 3.2, respectively.

In this chapter, I defend the interpretations of quality and status used in EEMM. I raise a worry about the quality interpretation in §5.2, namely, that it conflicts with the current interpretation of 'quality of evidence' used by the GRADE framework, which is put solely in terms of degrees of confidence. In §5.3 I appeal to the concepts of the weight and balance of evidence to justify the need for separate degrees and stability of confidence in evidence evaluation. While I find that this motivates the interpretation of status, it is not enough to motivate the interpretation of quality found in EEMM. To motivate the quality interpretation, I argue in §5.4 that criteria by which quality is assessed affect the weight of evidence. As weight of evidence is reflected in the stability of confidence, the EEMM interpretation of quality of evidence is correct. I finish by considering the nature of the relation between quality of evidence and weight of evidence (§5.5). I conclude that quality of evidence assessments are nothing more than weight of evidence assessments. The claims made in this chapter defend concepts central to evaluating evidence on EEMM, and propose an alteration to the way 'quality of evidence' is currently conceptualised in the GRADE framework.

5.2 The interpretation of quality of evidence: a problem

One can question whether the interpretations of the evaluative judgements are correct. The first interpretation to question is that of 'quality of evidence'. The reason it should be questioned is that it runs counter to the current interpretation of 'quality of evidence' used by the GRADE framework for evaluating evidence from clinical studies. The current GRADE interpretation of quality is that it "reflects the extent of our confidence that the estimate of effect is correct" (Balshem et al., 2011, p.404), where estimate of effect means a pooled estimate for the evidence base as a whole. So unlike EEMM, a quality assessment on GRADE does not factor in the stability of one's confidence. Compounding this problem further is the fact that GRADE changed its interpretation in 2011 from one that did make explicit reference to stability. For example, pre-2011, rating the quality of an evidence base as 'high' meant that "[f]urther research is very unlikely to change our confidence in the estimate of effect" (Guyatt et al., 2008, p.926). This way of interpreting levels of quality of evidence is exactly the same as EEMM. So GRADE have actively moved away from the interpretation used in EEMM. Why then the change, and is EEMM correct to retain the pre-2011 interpretation?

Before explaining the reasons for this change, some clarifications are in order. Firstly, while typically a GRADE evaluation results in confidence in an interval estimate that takes the form of a confidence interval, it is possible for a point estimate (e.g., the sample mean $\bar{x}$) to be used as well. Secondly, in more recent versions, the output of a GRADE evaluation is a rating of *certainty in the evidence* (Hultcrantz et al., 2017; Tikkinen et al., 2018). This new concept is however only a change in terminology, rather than a substantive change in the nature of the evaluation process. The features of evidence that contribute to the level of quality and certainty are the same. Moreover, when rating one's certainty in the evidence, one is rating one's confidence in the correctness of effect estimates. So one is using the criteria that affect ratings of quality of evidence to affect certainty of evidence, and both 'quality' and 'certainty' are interpreted in terms of degree of confidence. Therefore, we can still compare GRADE to EEMM as a method for evaluating the quality of evidence. Finally, talk of estimates of effect can be translated to talk of causal claims. For example, a common measurement of effect size is the risk difference (RD), which outputs a number in the interval [-1,1]. RD can be used to quantify differences in observed outcomes between the intervention and control groups in a clinical trial. Strength of the association between intervention and outcome can also be determined through computation of RD. Roughly, the closer the value of RD to 1, the stronger the association. If all we want to know is whether there is *some* causal relation between intervention and outcome, then RD > 0 will suffice. So to say that we are confident that an estimate of effect is correct is to also say that we are confident that there is or is not a causal relation (the role of effect size information in evidence evaluation on both EEMM and GRADE will be discussed further in Chapter 9).

5.2.1 The change from stability to degree of confidence

GRADE changed its definition of quality because “there are many situations in which we cannot expect higher quality evidence to be forthcoming” (Balshem et al., 2011, p.404). It might be the case that practical or ethical considerations preclude the conduct of further studies. The GRADE working group thought this posed a problem to evaluating evidence based on how likely future evidence, which would be impossible to obtain, would change one’s confidence. Parkkinen et al. (2018b) acknowledge both that their quality interpretation accords with pre-2011 GRADE, and that the GRADE framework has moved away from using this interpretation. However, they argue that the pre-2011 definition *is* legitimate. While further evidence may not be forthcoming, what matters is instead the *in principle possibility* of obtaining such evidence. This matters because in some cases we may have evidence that our degree of confidence in a claim should change, but to an otherwise indeterminate extent. Parkkinen et al. (2018b, p.26) offers an example to illustrate this, where current evidence warrants holding 75% confidence in a claim, but further evidence may warrant a 25% change in confidence. However, there is no indication of the direction of this change. One’s confidence on the future evidence could be 50% or 100%. All we have in such cases is an in principle possibility of change in degree of confidence, as we do not know what the value will change to. This is more easily captured by holding both degrees and stability of confidence. One holds a 75% confidence in the claim plus some measure of stability equivalent to a 25% swing in either direction. This would be instead of holding either 50%, or 100% confidence, neither of which is justified on the evidence.

A proponent of post-2011 GRADE might accept this argument. Indeed, in the paper that introduces the new quality definition, the GRADE working group note that “the prior characterization of quality [provides] an alternative under circumstances when obtaining new compelling evidence is plausible” (Balshem et al., 2011, p.404). This implies an acceptance of reasoning about the in principle possibility of obtaining future evidence, as such reasoning is required to make judgements about cases where it is possible to obtain future evidence. The 'prior characterisation' does not however find its way into later versions of GRADE, versions where GRADE evaluations exclusively rate degrees of confidence in effect estimates. Evaluators who have used GRADE will thus be familiar with the degree of confidence approach to quality. For example, GRADE was used to evaluate the evidence that contributed to a recent meta-analysis on the effectiveness of anti-depressants (Cipriani et al., 2018). This evaluation found the quality of evidence to be low-to-moderate, and what was rated was the extent of evaluator confidence in effect estimates. Moreover, it is practically possible to obtain more evidence on the effectiveness of anti-depressants, so the evaluators could have used the stability of confidence approach to evaluation. The fact that they did not, indicates that users of GRADE are currently working within the framework that employs degrees, not stability, of confidence.

There is thus a clash between how GRADE and EEMM interpret 'quality of evidence'. Moreover, EEMM offers only a hypothetical example to illustrate the claim that the pre-2011 definition suffices. This may not be enough to conclusively counter the practical examples in which GRADE evaluations proceed along post-2011 lines. Equally, neither have GRADE done the requisite conceptual work to justify an interpretation put solely in terms of degree of confidence. Therefore, it remains to be seen whether the pre-2011 or post-2011 GRADE interpretation is correct. In the remainder of this chapter I develop an argument in favour of the pre-2011 interpretation.

5.3 Weight and balance of evidence

Deciding between GRADE's pre-2011 and post-2011 interpretations of quality means deciding between whether levels of quality are interpreted in terms of either or both of the stability or degree of confidence in claims. A first step in making the argument for the claim that pre-2011 GRADE is correct is identifying why one should hold separate degrees and stability of confidence. Maintaining a strict separation between these aspects of confidence motivates the EEMM interpretation of statuses, and one reason to hold such a separation appeals to the concepts of weight and balance of evidence.

The weight and balance of evidence are two dimensions along which evidence can be strong. The first use of this distinction is attributed by Joyce (2005) to Keynes (1921). Importantly, the weight and balance of evidence are reflected in the stability and balance of confidence, respectively. The balance of the evidence makes a difference to what it is reasonable to believe. This could be through frequency data, as for example in a coin toss where 5 heads out of 10 flips of a coin indicates that, on the balance of the evidence, one should believe that the coin is fair. The particular balance of the evidence does not tell us anything about how substantial it is. 10 flips of a coin may not be enough to decisively conclude that the coin is fair. What we need is to flip the coin more to acquire more evidence. This increases the "gross amount of relevant data" (Joyce, 2005, p.154), but "may or may not make a difference...to what it is reasonable to believe" (Kelly, 2008, p.934). What this increase in the weight of evidence affects is instead "what it is reasonable to believe when additional evidence is acquired" (Kelly, 2008, p.935). How stable our beliefs are corresponds to the extent to which it is reasonable to believe something different when additional evidence is acquired. Note that even stable beliefs based on substantial evidence can change on additional evidence. But this evidence would have to be quite exceptional, e.g., that the methods by which a substantial body of evidence for a belief is generated, are subsequently found to be systematically biased.

Joyce (2005) argues that the weight and balance of evidence are reflected in the probabilistic profile of beliefs. He does so in the context of arguing for a specific Bayesian epistemology, details of which I expand on in §7.2.3. But he uses a basic understanding of probabilities as measuring degrees of belief, and those probabilities are assigned by direct inference from frequency data: for attribute A of event $\mathcal{E}$, if the frequency with which A occurs is 30 and the total number of $\mathcal{E}$ is 100, then the frequency of A is $\frac{30}{100}$, and $P(A) = 0.3$. His arguments show both how the distinction is reflected in one's confidence and why it is important to do so. An urn example is used to illustrate his point. Two people, Emily and Jacob, draw green and blue balls from four urns and use the frequency of blue and green balls to estimate probabilities for whether the next ball from an urn will be blue (B). Jacob draws 5 blue and 0 green (total of 5 balls), and Emily draws 17 blue and 3 green (total of 20 balls). By the balance of evidence, Jacob is more confident that the next ball is blue ($P(B|E) = 5/5 = 1$) than Emily ($P(B|E) = 17/20 = 0.85$). However, Emily's evidence is more weighty, as she has 20 draws to Jacob's 5. Joyce argues that this means she has a "more settled picture of the situation" (2005, p.161). This is because any further draws of balls from the urn will change Jacob's probability to a greater extent than they would Emily's. Five green balls will shift Jacob's probability from 1 to 0.5, but will shift Emily's probabilities from 0.85 to 0.68. Emily's confidence (measured by her probabilities) is more stable than Jacob's and this is a result of having weightier evidence.

Importantly, acknowledging this distinction allows a better judgement on the strength of evidence. In the case above, we can say that Emily's evidence is better than Jacob's. While Jacob's evidence is more strongly in favour of the hypothesis, his high confidence in the truth of B is arguably mistaken. And we judge that he is mistaken not on the basis of his degree of confidence in B (=1), but on the basis that his evidence is not substantial. We therefore expect his degree of confidence to change when more evidence is obtained. The distinction is also important for medicine. Consider a causal claim such as 'drug D causes outcome O'. Suppose that an RCT is used to test the efficacy of this intervention and observes a difference in O between intervention and control group by a factor of 80%. The evidence therefore strongly favours the hypothesis that D causes O, and we would be highly confident in the causal claim. If the evidence base consists of only this one study, then the evidence is not weighty. In this sense the evidence base should not strongly support the claim. But it is difficult to put this solely in terms of degree of confidence. Knowing only that we have one study does not tell us whether, if we were to have more studies, we would find a 50% improvement in O instead of 80%. So the balance of our evidence does not change and the evidence is still strongly in favour of the hypothesis. Having a notion of stability of confidence aids us here. Ignoring this aspect of our evidence would mean we think our evidence is stronger than it really is. This is important for decision making. High confidence may have led to approving this drug. Identifying that this confidence is unstable should instead lead us to doubt whether the drug should be approved. If we did not account for the weight of evidence, this drug would have been wrongly approved. The reasoning behind this example is the same as that in the example found in Parkkinen et al. (2018b, p.26) (see §5.2). While we don't know whether more evidence would result in a change to some specific result, it is likely that there would be a change. So our confidence in the results is unstable. This is why we should hold separate degrees and stability of confidence, why the balance of evidence affects what degree of confidence one holds, and why the weight of evidence affects the stability of that confidence.

One might think that in both the urn and medical case one should not have held high confidence on such misleading evidence. Instead, one might make the stipulation that only when the weight of the evidence is high should one hold high confidence. Call this the *HighHigh* view. When Jacob draws 5/5 blue balls, he sets his confidence to 1. Given that evidence should plausibly be substantial, as well as in favour or against a hypothesis, holding such high confidence should not be warranted on such insubstantial evidence. The same can be said for clinical trials. When a new drug is tested in a trial, and suggests that the drug works, claims about effectiveness are typically tempered by acknowledging that more research is needed. Importantly for the argument of this chapter, a strict relation between weight of evidence and stability of confidence is rejected on HighHigh.

I think however that the way I utilise the weight / balance distinction in this chapter is better motivated. To aid clarity in the remainder of this section, I will call the view I support the *HighLow* view, as on this view it is possible to hold high confidence on low weight of evidence. There are two reasons to prefer the HighLow view. One is that it can account for whether evidence is misleading, through the stability of confidence, while also keeping more in line with what the evidence says it is reasonable to believe. The second is that it does not face a conceptual problem that besets the HighHigh view, namely, the threshold over which one has sufficiently weighty evidence to hold high confidence. Both reasons rely on acknowledging two things about the HighHigh view. First, what happens on the HighHigh view when insubstantial evidence strongly favours a hypothesis? Plausibly, one should hold moderate degrees of belief. For example, consider again the hypothesis B, that the next ball drawn from an urn is blue, where the colours and total numbers of balls in the urn are unknown. Suppose that the evidence for B is that 5/5 balls drawn are blue. On the HighHigh view, one should not hold confidence equivalent to 1, because of the insubstantial nature of the evidence. Instead, holding degree of confidence equivalent to $\mathrm{P}(B) = 0.5$ is better motivated. One might plump for some other degree of confidence, less or greater than 0.5, but plausibly no greater than 0.8. However, the exact value assigned to B is not important for my claims as long as it is interpreted as moderate. Next, we must acknowledge that even on the HighHigh view, the stability of confidence is still important. One cannot consider $\mathrm{P}(B) = 0.5$ to be the final degree of confidence in B, as one has to recognise what values $\mathrm{P}(B)$ may take once we require more evidence. HighHigh might say that $\mathrm{P}(B) = 0.5$, and $\mathrm{P}(B) \in (0,1]$ are possible given that it is possible that all of the balls in the urn are blue, and there is at least 5 blue balls. The HighLow view would say that $\mathrm{P}(B) = 1$, and $\mathrm{P}(B) \in (0,1]$ are possible (for the same reasons as in the HighHigh case). In both cases we are saying that there is a range of possible degrees of confidence to hold, from within which we select a value to hold as our actual degree of confidence.

One might think that the representation of confidence on HighLow is better motivated because it is informed by the information provided by frequency data. Relying on the role of frequency data to make this case means identifying why it is important for it to contribute to the balance of evidence. I set out above how the views only differ on how they represent the balance of the evidence. And the problem with how the HighHigh view does this, is that it assigns actual degrees of confidence based on what values are merely possible. On the other hand HighLow assigns them on the basis of the frequency data, which is the only actual indication of what the true proportion of blue balls is in the urn. Moreover, the HighLow view includes a representation of stability of confidence, which can account for the misleading nature of the evidence. Crucially, this is no different than the representation of stability on the HighHigh view. For those reasons, one might think that HighLow is better motivated as it keeps more in line with the evidence we do have than HighHigh, and can account for misleading data. However, the HighHigh view does not totally ignore the frequency data. The information that 5/5 balls have been drawn blue restricts the possible degrees of confidence that might be held to (0,1]. So it is to some extent in line with the evidence, and a proponent of the HighHigh view might not be convinced by this argument.

Another reason to prefer the HighLow view is that HighHigh suffers from a conceptual problem. A consequence of the HighHigh view is that there must be some threshold over which the weight of evidence is deemed sufficient for one to hold high confidence. Determining such a threshold may be difficult. In the case used thus far, it is fairly easy to say that 5 total draws is not sufficient to denote high weight. However, consider the same urn drawing experiment, but at a stage where 1000 balls have been drawn, of which 900 were blue. We might be inclined to say this evidence has high weight. But is it high enough to say we should hold a degree of confidence equivalent to $\mathrm{P}(B) = 0.9$? Even if we knew that there are 2000 balls in the urn, determining a point at which we have high confidence might be arbitrary. Moreover, whether we are highly confident in B would depend on the threshold we choose for what constitutes high weight, tying our confidence to the threshold rather than to the evidence. On the other hand, HighLow has a fairly straightforward way of representing confidence in this case: $\mathrm{P}(B) = 0.9$, and the range of values possible is [0.45,0.95]; the lower bound is calculated on the assumption that no more blue balls are drawn, and the upper bound on the assumption that only more blue balls are drawn. Because we represent the weight of evidence using stability of confidence, the evidence is not misleading, and we keep in line with the evidence we do have for how confident we should be in B. Moreover, we do not have to worry about a threshold over which the weight of evidence is sufficient to hold high degrees of confidence. So HighLow is not burdened by any additional conceptual problem, and is better motivated than HighHigh.

On the view set out above, the core idea behind holding separate notions of stability and degree of confidence is that there is more to whether evidence is strong than how strongly it favours an hypothesis. This motivates having an interpretation of status in terms of stability *and* degree of confidence. It does not however force the view that quality of evidence should be interpreted in terms of stability of confidence. I have only set out why the *weight* of evidence should be reflected in the stability of confidence. In the next section I give reasons for thinking that the quality of evidence is at least part of what contributes to the weight of evidence.

5.4 Quality and weight of evidence

In this section I appeal to philosophical accounts of evidence, and the practice of evidence assessment, to support the claim that 'quality of evidence' should be interpreted in terms of the stability of confidence. Philosophy and practice both show that there is a link between quality and weight of evidence. I then delineate the nature of this link. The basic idea is that there is something about evidence that gives it a level of quality, and this is what is evaluated by *quality criteria* (e.g., those found in Parkkinen et al. (2018b, pp.80-82), and the moderating domains of GRADE (§1.3)). There is also something about evidence that gives it weight. I first analyse the *abstract properties* of evidence that give it weight. Those properties are not particular to methods used to obtain evidence, but apply in general to evidence from all methods. *Features* of evidence are particular to implementations of methods; they are typically what a quality assessment looks for. I will show how the particular features of evidence by which quality is determined have the properties of evidence that give it weight. Given that the weight of evidence is reflected in the stability of confidence, what counts towards the quality of evidence should be as well. Therefore, the pre-2011 GRADE interpretation of 'quality of evidence' is correct.

5.4.1 What gives evidence its weight?

Relative to some hypothesis, evidence is weighty or has balance because of some property of the evidence. As I noted in §5.3, it seems like the gross amount of information the evidence contains contributes solely to the weight of evidence. For example, 1000 coin flips provides greater weight of evidence than does 5 coin flips. The fact alone that the coin is flipped 1000 times contributes nothing to the balance of the evidence as it tells us nothing about the proportion of heads and so what degree of confidence in heads we should hold. But our confidence on the evidence base that contains 1000 flips is more stable than the one that consists of 5 coin flips. A similar example from statistics is the link between precision of interval estimates and sample sizes. In the confidence interval approach to estimation, the larger the sample size, the more narrow the interval tends to be. A large sample size will include more data, so is more informative than a small sample size. The width of a confidence interval need not be a representation of the stability of confidence. But this example does show how considerations of informativeness are linked to considerations of precision.

I distinguish here between two kinds of informativeness. *Gross informativeness* concerns the amount of evidence/data that is available. For example, gross informativeness can be effected by the amount of measurements made in one experiment, the number of experiments in a pooled evidence base, or the number of testifiers in the reporting of an event. *Net informativeness* concerns the informativeness of the evidence base once one takes into account any inadequacies in how that evidence was generated and reasoned with. For example, the total number of times a coin is flipped may be 1000. This evidence would have a higher gross informativeness than if the coin were flipped 5 times. But we might then find out that the coin flipper has a tendency to misreport whether the coin lands heads or tails. Gross informativeness stays the same, but this inadequacy in the generation of the evidence would reduce the informativeness of the evidence as a whole. Because of this, the net informativeness of the evidence, and so also its weight, is low. However, the notion of 'inadequacies' used in the formulation of the notion of 'net informativeness' is quite vague. What makes some feature of the evidence an inadequacy? And what is it about inadequacies that reduces the informativeness of the evidence?

My answer to these questions is that inadequacies pose alternative explanations, and evidence bases lose informativeness when one can explain results in terms other than that the hypothesis is true. Consequently, one's confidence in that hypothesis becomes unstable. I will develop this proposal by first acknowledging an initial problem, namely, that explanatory information is typically considered unable to contribute to probabilistic representations of confidence (for summaries of this issue see Okasha (2002) and Bird (2018)). For instance, Roche and Sober (2013) consider the evidence for the causal hypothesis C = 'smoking causes lung cancer'. We first obtain frequency data on the incidence of lung cancer L and smoking S in a test population. To be specific, the frequency data provides evidence about the *objective chance* of C, where the objective chance is the 'true', but unknown, probability of C. The frequency data is used to calculate a posterior probability $P(L|S)$, which denotes the probability of getting lung cancer given that someone smokes. On the terminology introduced in §1.1, this probability would be used to determine whether L and S are appropriately correlated; i.e., when $P(L|S) \neq P(L)$. Roche and Sober are not working within this framework, but below I will talk as if determining $P(L|S) \neq P(L)$ determines that S causes L. Roche and Sober argue that learning that smoking explains cancer would not alter the value of the posterior probability, as the explanation tells us nothing more about the objective chance of getting cancer, given that one smokes, than the frequency data does. Contrary to this position, McCain and Poston (2014) argue that explanations do contribute to the probabilistic profile of one's beliefs. They agree with Roche and Sober that explanation does not effect the *value* the posterior probability takes. But they object to the view that explanations do not affect probabilities at all. Instead, the posterior probability is *stabilised* by learning this explanation.

To illustrate this claim, McCain and Poston (2014) present the 'x-sphere example'. In this variation on the classic urn sampling example, an urn is filled with 1000 x-spheres. Sally and Tom observe the random drawing, without replacement, of 10 x-spheres from the urn: 5 blue and 5 red. The x-spheres are then put back in the urn. Given the data, and so on the balance of the evidence, both Sally and Tom should assign P(blue) = 0.5. Another 10 x-spheres are drawn at random and they are all blue. On the balance of the evidence Tom assigns P(blue)=0.75. However, Sally has explanatory information that Tom does not: due to their atomic structure, blue and red x-spheres must be stored in equal numbers otherwise the atoms of all the x-spheres will spontaneously decay causing a massive explosion. As they did not all explode after observing 10 successive blues, Sally keeps her confidence at P(blue)=0.5. They interpret this as Tom's confidence being volatile, while Sally's is more resilient in the light of new information. Crucially, the explanation does not contribute to the value of the assigned probability, but rather to its stability by fixing the probability at 0.5. So explanations can contribute to the *stability* of confidence, by providing information about whether the frequency data is correct about the value of the objective chance.

That explanations can stabilise or destabilise confidence is not however sufficient to conclude that they also contribute to the weight of evidence. It is still possible that confidence is stabilised by some other route than through increasing the weight of evidence. For explanations to affect stability of confidence by varying the weight of evidence, they must play a role in reducing the informativeness of the evidence. One way they can play this role is to see inadequacies as posing alternative explanations, presence of which reduces the informativeness of an evidence base. I argued in Chapter 2 that the methodology of ruling in or ruling out alternative explanations is central to evaluating causality. It is plausible that this same reasoning process is central to evaluating evidence as well. As seen in the x-sphere example, the explanation that there can only be equal amounts of x-spheres rules out all competing explanations. The converse would be true if Sally had some evidence that supported an alternative explanation, e.g., that 20% of all x-spheres do not explode when the proportion of blue to red is within the range 25/75 to 75/25. This information rules in the objective chance hypotheses within this range, while not picking out which one is correct. Sally cannot change her confidence to another value on the basis of this information but it does mean her beliefs are unstable. In much the same way, inadequacies in the generation and reporting of evidence can lead to instability of confidence. Suppose that the evidence that x-spheres can only be stored in equal numbers was obtained from one study only. This would typically be an inadequacy, as one study would likely not be large enough to collect a quantity of data requisite for ruling out the possibility of random error. Such an inadequacy rules in an alternative 'random error' explanation of the results that led to the conclusion that x-spheres must be stored in equal numbers. Moreover, this alternative explanation does not tell us what the proportion should be. Sally thus does not change the degree of her confidence, but instead its stability. And this is because we have less information about the objective chance of plucking a blue x-sphere from the urn than we otherwise would have, had we had an evidence base consisting of many studies. The presence of alternative explanations thus reduces the weight of the evidence, through reductions in informativeness, rather than by altering its balance.

To sum up, the weight of evidence should be understood as measuring the extent of the informativeness of the evidence. More precisely, it is the net informativeness of evidence, where net informativeness is worked out by consideration of explanations of the results in terms other than that the hypothesis under consideration is true. These *alternative explanations* are posed by finding inadequacies in the design, implementation and analysis of a study. Where alternative explanations are present, the evidence is less weighty than it otherwise would be were those alternative explanations not available. These explanations should indicate that, if future evidence is obtained which rules them out, then one's confidence in the claim or estimate should also change. In all cases one is becoming less confident that the true chance is equal to the value determined by the evidence. And as we do not have evidence to pick out any of the alternatives as the true chance, our confidence becomes unstable rather than being assigned a new value. This sets out the properties of evidence that give it weight. In the following, a particular feature of evidence that contributes to the weight of evidence will have the property that it contributes to the extent of net informativeness of the evidence.

5.4.2 What gives evidence its level of quality?

To make the case that the quality of evidence is reflected in stability of confidence, I argue here that what counts towards the quality of evidence is part of what contributes to its weight. To do so, I appeal to support in the philosophy of science literature and to the practice of evidence evaluation on the GRADE framework. In both cases, the features of evidence that contribute to its quality can be made sense of in the net informativeness framework I argued for in the previous section. This means they have the properties of evidence that give it weight. Given that weight of evidence is reflected in the stability of confidence, 'quality of evidence' should therefore be interpreted in terms of stability of confidence.

Characterisations of 'quality of evidence'

One characterisation of quality of evidence is in terms of *freedom from error* in individual studies, and finds support in the clinical and philosophical literature. A prominent clinical research textbook, '*Clinical Epidemiology: The Essentials*', instructs (systematic) reviewers that "individual elements of quality... present in the studies" (Fletcher et al., 2012, p.213) should be included in a review. The link between quality and individual studies is echoed in the critical literature (see, e.g., Grossman and Mackenzie (2005, p.523) and La Caze (2009, p.523)), and this link seems uncontroversial. In clinical epidemiology, talk of quality is often put in terms of internal validity (Balshem et al., 2011), where internal validity means "how likely are the results of the study to be true for the participants involved" (La Caze, 2009, p.519). Internal validity is determined by consideration of "how well the design, data collection, and analyses are carried out" (Fletcher et al., 2012, p.11). La Caze (2009) interprets this as meaning internal validity is about minimizing a range of errors during the design and implementation of studies.

Stegenga (2013, 2018) makes this characterisation more precise and I will focus on it from here on. He defines 'quality' as "the extent to which the design, conduct, analysis, and report of a study minimizes potential bias and error" (Stegenga, 2018, p.99). Support for this characterisation is levied by appeal to case studies of experimental work in biology (Stegenga, 2013). For example, in the research that discovered that nucleic acids in the form of DNA carried 'genes', when trying to determine whether it was in fact nucleic acids that were producing the effect of gene transfer in bacteria, an important consideration was whether all sources of contamination were ruled out. Stegenga interprets this as an assessment of freedom of methodological error, giving the evidence a high level of quality. Stegenga's characterisation proceeds from an analysis into different traditions of philosophical accounts of evidence. The first step in this analysis is to make a distinction between different traditions of philosophical accounts of evidence. Accounts in the *signs of success* tradition "[describe] what is achieved once one has reliable evidence" (Stegenga, 2013, p.983). Signs of success accounts often involve some use of the probability calculus to describe what good evidence achieves. Many such accounts hold that if $P(H|E) > P(H)$, i.e., the probability of the hypothesis H conditional on some evidence E is greater than the probability of H alone, then E is evidence for H. Different accounts then might say something about what sort of posterior probability indicates that E is evidence for H

(e.g., Achinstein (2001, ch.6) argues that E is evidence for H when $P(H|E) > 0.5$). This is contrasted with accounts in the *conditions of success* tradition which are "prior to and richer than the signs of success tradition" (Stegenga, 2013, p.982). This is because they describe important features of evidence that would lead to a rise in the probability of H conditional on E. Stegenga then distinguishes further between different conditions of success, which are divided into methodological and evidential features.

Quality is one methodological feature, as are *relevance* and *transparency.* Studies do not provide evidence in isolation, but instead always in relation to some hypothesis. Evidence may be positive for one hypothesis but irrelevant to another. For example, if a study focuses on a narrow set of objects then its evidence is not relevant to a wider hypothesis. The relative transparency of an evidence base concerns how easy it is to assess a method for either of the two other methodological conditions. A method may be completely novel and so all its possible sources of error hard to define; such a method would be opaque and hard to assess. Evidential features are particularly important when a method is opaque, as they can be used to indicate that the evidence is or is not reliable. Stegenga identifies three evidential features. One, *patterns*, refers to there being identifiable and consistent patterns in the data generated by the studies under assessment. That a pattern exists gives some reason to believe that what is being detected is in fact a real phenomenon. A second, *concordance*, refers to obtaining coherent and consistent results across an evidence base that consists of studies utilising both similar and different methods. If two different experimental techniques or methods provide evidence that points in the same direction, then there is good reason to believe that the evidence concerns something real. Finally, *believability* means that the evidence must fit with current physical possibility. If our evidence contradicts the laws of gravity we would tend not to believe that the laws were false, but instead that there is something wrong with the evidence.

I claim that the conditions of success that Stegenga identifies, contribute to the weight of evidence. Each 'condition' poses an alternative explanation, thus reducing the net informativeness of the evidence base. Errors all pose alternative explanations for the results. If an experiment does not properly implement controls for a particular kind of bias, then it is possible that the bias is responsible for the results. Evidence bases with errors are less informative than those without errors. If an evidence base is not relevant to an hypothesis, then it is less informative about that hypothesis than it otherwise would have been. When a method is novel, making it hard to assess evidence obtained from it, we lack information about how the evidence is generated. An alternative explanation of the results other than that the hypothesis is true would focus on what aspects of the method are opaque, e.g., what the controls are, what the standard measurement errors are. Identifying patterns or concordance rules out alternative explanations, making one's confidence more stable. The absence of patterns, or presence of significant discordance, would introduce alternative explanations, thus leading to unstable confidence. They do so by giving us reason to believe that the extent of our confidence in a claim is not correct. So, the conditions of success are all features of the evidence that contribute to the weight of evidence. Quality is a condition of success and so is part of what contributes to the overall weight of evidence. Therefore, on the basis of philosophical accounts of evidence, quality of evidence affects stability of confidence, and the pre-2011 GRADE interpretation of 'quality of evidence' is correct.

Evaluation of the quality of evidence in practice

This view also finds support in the *practice* of using the GRADE framework to carry out an evidence evaluation. In the switch from pre- to post-2011 definitions of quality, what GRADE call the 5 key moderating domains, broad categories of criteria within which are the more precise methodological details by which evidence is ultimately assessed, have remained the same (see Table 1.1 on page 22). The five domains are: *Risk of Bias*; *Inconsistency*; *Imprecision*; *Relevance*; *Publication Bias*. Where domains have been altered, it is by refining, rather than adding or removing, criteria. Thus there is a high degree of consistency between the way the quality of evidence is evaluated between pre- and post-2011 GRADE. I claim that quality is evaluated by criteria that all contribute to the net informativeness of the evidence. The criteria identify inadequacies in the generation of the evidence and/or the reasoning carried out on the basis of that evidence. Crucially, all inadequacies pose alternative explanations for the estimate of effect than that it is the result of a real causal relation with that magnitude. So, although the definition of quality has changed, the unchanged process of evaluating quality supports a definition in terms of stability of confidence. To support this claim I next make sense of each domain in terms of my net informativeness framework.

It is straightforward to make sense of risk of bias and publication bias in the 'alternative explanations' framework I have proposed. If any of the biases are present then they are potential alternative explanations of the estimate, reducing the weight of evidence. Alternative 'bias' explanations *may* indicate a direction in which the true value should be. It may be known whether a bias causes an over- or under-estimate of the effect. One way to obtain this knowledge is through carrying out meta-epidemiological studies to measure the effect of biases. For example, Saltaji et al. (2018) found that improper allocation concealment in studies on oral health interventions resulted in effect sizes 0.15 larger than in studies that properly concealed allocation. This inadequacy results in unbalanced causal factors across experimental arms in clinical studies, and the bias caused by it is known as *selection bias*. One might use such information, where it exists, to affect changes in degrees of confidence. But the direction and magnitude of changes in effect as a result of bias in any one study is typically not known. In such cases it is more appropriate for the stability of one's confidence to be effected.

An evidence base is *inconsistent* when there is an unacceptable degree of conflict between the results of the studies it consists of, e.g., 50% are in favour, 50% against, a hypothesis. In a large evidence base, due to random error, it is expected that a specifiably small proportion of studies will conflict with the rest. It is also the case that in small evidence bases, heterogeneity can go undetected. To quantify the extent of unacceptable inconsistency, GRADE propose using the I^2 score, which "describes the percentage of total variation across studies that is due to heterogeneity rather than chance" (Higgins et al., 2003, p.558). Finding an evidence base to be unacceptably inconsistent reduces its informativeness. Typically, the 'estimate of effect' will be calculated by pooling the estimates from each study. An unacceptable degree of inconsistency indicates that only some, or even none, of those studies obtained 'correct' estimates. An alternative explanation of the pooled estimate, other than that it is correct, is that it is a result of studies that do not accurately represent the underlying effect, if there is one. In an inconsistent evidence base it is hard to know which studies demonstrate incorrect estimates, so it is hard to pinpoint what the effect should be. But the entire range of values obtained in these studies indicates the change that is possible, were those studies to be conducted better. So, again, stability in one's confidence is effected.

Imprecision as a domain is difficult to fit into the way evidence is assessed in the rest of the framework. The vagueness of this imprecision domain has been criticised by Antilla et al. (2016), to which the GRADE working group have responded with clarifications of it (Schünemann, 2016). On a basic understanding, imprecise evidence bases are those where the interval estimates of effect are too wide. However, GRADE's most recent iteration treats this domain's effects on confidence differently to the way the other domains effect confidence. The importance of the width of an interval seems now to be relative to decision making, rather than an assessment of whether the estimate of effect is correct. Evidence does badly on the imprecision domain when the interval crosses some pre-specified threshold (the delta (δ) level) at which an intervention is deemed to display a beneficial effect (an idea which is developed in more detail in Chapter 9). This threshold will be context dependent, and even a very wide interval, if it does not cross δ, will not be considered to do badly on the imprecision domain. It is not clear whether this element of the imprecision domain still has a place in the standard GRADE assessment of the correctness of estimates of effect, as it now fits more into the 'strength of recommendations' part of the framework. This part concerns whether evidence is adequate to support recommending the intervention under assessment, rather than whether an estimate is correct.

GRADE do however still allow for a rough assessment of the width of interval estimates. Wide interval estimates indicate that there was a high degree of variability in individual measurements of effect size. They are often linked to sub-optimal sample sizes. When sample sizes are sub-optimal, there is a greater chance that any observed effect is due to random error. In other words, that there is no ‘true’ effect. A wide interval will thus rule in an alternative explanation for the observed effect: random error. It is also the case that with a wide interval, the estimate makes probable many parameter values that are not the ‘true’ value of the effect. When imprecision is interpreted in this way, it contributes to the weight of evidence by introducing the random error explanation and/or opening up the space of probable alternative parameter values. There is a final clarification that must be made about even this interpretation of the imprecision domain. The most recent forms of GRADE talk about treating the interval estimate as the estimate we are supposed to hold confidence in (Hultcrantz et al., 2017). If this is the case, then the width of the estimate will be irrelevant to whether the true effect is in the estimate or not. So, to be able to evaluate the imprecision domain, GRADE should be interpreted as evaluating whether some point estimate, taken from within the interval estimate, is the true effect. In only this sense does a wide interval rule in many parameter values as probable alternatives for the true value.

The *relevance* domain is broadly similar to Stegenga’s relevance condition of success. On the GRADE framework, evidence bases are assessed relative to a question, and that question will concern the specific outcomes observed from testing specific interventions on equally specific populations. If an evidence base tests something other than this, then it does badly on the relevance domain. For example, if the hypothesis H is ‘intervention I causes outcome O in sub-population p’, and evidence base E demonstrates that in sub-population q, I causes O, then E is not relevant to H. An alternative explanation for the estimate of effect is that it is an estimate for the effect in sub-population q. Finding inadequacies in the evidence base on this domain, opens up a space of estimates that are possible were the evidence base not to include those inadequacies. As no values are picked out, our evidence becomes unstable.

Philosophical accounts of evidence support the idea that the level of quality of evidence is relevant to assessing the weight of evidence. The practice of evidence evaluation, in particular the GRADE framework, also supports this idea. Domains that evaluate quality do so through contribution to the weight of evidence. As weight of evidence is reflected in stability of confidence, the pre-2011 interpretation of quality, and by association that of EEMM's, is correct.

5.5 The relation between quality and weight of evidence

One might think that there is a problem with appealing to both Stegenga's conditions of success and GRADE practice as they clash on how 'quality of evidence' is characterised. The key point of discordance is that some of the GRADE quality criteria fit into conditions of success other than the quality condition. For example, relevance and inconsistency are features of evidence that matter for an assessment of quality on the GRADE framework, but are distinguished from the quality 'condition of success' on Stegenga's framework. Additionally, on GRADE, some features of evidence that fall under the patterns 'condition of success' are used to up-rate the quality of evidence. For example, if the evidence displays a 'dose-response gradient' then one can rate-up the level of quality of evidence obtained from non-randomised studies. A dose-response gradient involves observing that the magnitude of the effect increases as the dose administered increases. This makes the observed effect more likely to be the true effect even when the study design cannot rule out confounding. For the same reasons, dose-response gradients are considered by Stegenga to be one kind of pattern to look for when assessing evidence. However, 'patterns' as a condition of success is distinguished from 'quality'. So there are multiple points of discordance between Stegenga's characterisation of 'quality', and the practice of evaluating the quality of evidence on the GRADE framework.

This discordance between Stegenga's distinctions and practical examples in evaluative frameworks is not confined to GRADE. Inconsistency, concordance, and patterns are all important for evaluating evidence in the part of the EEMM framework that deals with 'consistency of results' (introduced in §3.4.2 of this thesis; details can be found in Parkkinen et al. (2018b, pp.81-82)). A factor that EEMM, but not GRADE, considers relevant to the evaluation of individual studies is transparency. This makes sense as EEMM is concerned with the evaluation of mechanistic studies. The methods used in mechanistic studies can often be complex and highly novel, so transparency will be a key concern. Indeed, to support the transparency condition of success, Stegenga appeals to case studies from experimental biology, rather than clinical science. However, EEMM think transparency is pertinent to quality. There is therefore a discordance between Stegenga's characterisation of quality of evidence and how quality is assessed in practice.

However, this discordance does not pose a problem for my claim that quality is reflected in the stability of confidence. For Stegenga, quality is just one of many conditions of success, which I have argued contribute to the weight of evidence, whereas for GRADE it would appear that there is little more to assessing the quality of evidence than an assessment of the weight of evidence. Core to either view is that quality *in some way* affects the stability of one's confidence. As a consequence, the EEMM and pre-2011 GRADE interpretations of quality are correct, and the post-2011 GRADE interpretation is incomplete. However, it is still important to adjudicate on the correct way of characterising quality; as one part of weight of evidence, or nothing more than weight of evidence. As identified in §5.4.2, the understanding of quality as 'freedom from error' is not confined to Stegenga: it is also held in clinical science. Moreover, 'quality of evidence' is an important term in the field of evidence evaluaton. Stegenga (2018) argues that quality assessment is better than evaluation by evidence hierarchy because quality assessments are more informative. Evidence hierarchies "are rank-orderings of kinds of research methods according to the potential for those methods to suffer from systematic bias." (2018, p.71). The strength of evidence is then evaluated relative to the kind of method used to obtain it. For instance, I made reference to a very simple hierarchy of study-designs in §1.1, which is represented visually in fig. 1.4, on page 8. Stegenga (2018) argues that quality assessment is more informative than evaluation by hierarchy because of the centrality to the former of assessing by fine grained methodological details. Such frameworks evaluate individual studies for how well they have designed and implemented their methods, and analysed their results. Much more information about an evidence base can be obtained when implementation, and not only study-design, is considered. This claim is supported by the fact that when conducting meta-analyses one must evaluate the quality of evidence. Indeed, the meta-analysis of Cipriani et al. (2018) introduced in §5.2 utilised GRADE to assess the quality of the evidence that contributed to the study.

Obtaining clarity on the correct interpretation of 'quality of evidence' is thus important for matters of consistency. Methods throughout science need terms with consistent definitions and/or interpretations. This is so they can be applied consistently and correctly across different contexts. For example, in the field of statistical inference the terms *statistical significance* and *confidence interval* are often interpreted in variable ways. However, there are standard ways in which the concepts should be interpreted and how they are interpreted matters for statistical inference. This importance is demonstrated by the need to continually communicate and re-assert the correct interpretations, often by contrasting them with incorrect interpretations (see, e.g., Morey et al. (2016) and Greenland et al. (2016)). The concern is that incorrect interpretations of these concepts will lead to incorrect applications of them in statistical inference. Quality assessment frameworks are no different. Consistency and correctness of the term 'quality of evidence' is still important because incorrect interpretations may lead to the application of the wrong kind of assessment framework. Consider again Stegenga (2018)'s arguments for the claim that quality assessments are better than hierarchies. He appeals to Quality Assessment Tools (QATs) as examples of quality assessment frameworks. One could choose a QAT to assess the quality of evidence when conducting a meta-analysis. Alternatively, one could choose GRADE. Deciding which one to pick assumes that they are assessing the same thing. But, QATs use the term 'quality' in the more restrictive sense of 'freedom from error', as they typically only assess for risk of bias. On my view, QATs are not assessing all that is relevant to quality, whereas a framework like GRADE does: freedom from error is just one feature of evidence that contributes to its weight. So the correct characterisation of quality matters as it will help to determine whether frameworks like QATs are valid ways to assess evidence.

It is best to see quality assessments as weight of evidence assessments because they will result in more informative assessments than if quality is seen as 'freedom from error'. Informativeness is common ground for what counts as the best way to assess evidence. Stegenga is a proponent of the freedom from error view, and sees QATs as better than evidence hierarchies on the grounds that they are more informative. But on this count, assessment frameworks such as GRADE and EEMM are more informative than QATs, which only consider possible sources of bias and error. This is because weight of evidence is contributed to by all of the conditions of success, of which just one is 'quality of evidence' understood as 'freedom from error'. So an assessment of just 'freedom from error' would be less informative than an assessment by the weight of evidence. Characterising quality in this way would therefore be incorrect and it is best characterised in terms of assessing the weight of evidence. GRADE and EEMM are therefore valid frameworks for assessments of the quality of evidence, whereas QATs are not.

Stegenga might object to the use of GRADE as support for the argument that what counts towards the quality of evidence counts towards the weight of evidence, on the grounds that QATs *are* more informative than GRADE. Indeed, he levels some criticisms against the informativeness of GRADE. One criticism is that GRADE communicates the strength of evidence on an ordinal, rather than continuous scale. Stegenga argues that continuous scales are more informative than ordinal, and QATs rate evidence on continuous scales. This is however a misunderstanding, as although GRADE's quality rating system "represents discrete steps on an ordinal scale, it is helpful to view confidence in effect estimates as a continuum." (Guyatt et al., 2013, p.157). Confidence should be communicated on a continuum, according to GRADE, when judging threshold cases that could easily fall into one of two quality levels. Otherwise, ordinal levels are easier to communicate. His other criticism is that GRADE has a two stage process of first evaluating by one property, study design, and then by other properties, the quality criteria. This makes GRADE less informative as the first property heavily influences the final judgement. As Mercuri et al. (2018) argue, this feature of the framework means that a randomised study with a serious limitation will end up with a rating of quality that is similar to a non-randomised study with no limitations. So clinical decisions would be based on flawed evidence bases, but would ignore non-flawed ones. Mercuri et al. (2018)'s argument is a more complete development of the criticism Stegenga makes, so I opt to respond to it here.

Mercuri et al. (2018)'s argument depends on whether one thinks randomisation is a necessary technique for causal inference. As noted in §1.1, the supremacy of randomisation as a technique for justifying causal inferences has been challenged. Mercuri et al. (2018) support such critiques. GRADE should therefore not be justified in holding randomisation in such high esteem that a bad randomised study results in similar confidence as a good non-randomised study. However, Mercuri et al. (2018) ignore the way evaluative criteria on GRADE can both up-rate and down-rate quality levels. A well conducted non-randomised study *may* result in confidence in the effect equivalent to a randomised study with a limitation. But to then say that this rating is wrong, because of the critiques of randomisation, ignores the part of those same critiques that compare randomisation and *other* tools for ruling out confounding. The idea is that randomisation is just one technique for ruling out confounders, and that non-randomised studies can also employ their own techniques, e.g., matching of known confounders (§2.2.4). And on GRADE, the quality of evidence from non-randomised studies can be rated up if the studies employed techniques for ruling out confounding. There are also other ways the quality of evidence can be rated up, e.g., observing a dose-response curve. So it will not always be the case that a well conducted non-randomised study will result in confidence in the effect equivalent to a randomised study with a limitation, just so long as it also employs techniques, or obtained results, that help to rule out confounding. To maintain that GRADE's initial rating by study-design is wrong, Mercuri et al. (2018) would thus have to argue that a non-randomised study that did not do this should result in a higher quality of evidence rating than a randomised study with limitations. But in both cases, confounding is still a problem, so arguably the rating of quality of evidence should be roughly equivalent.

The initial ratings of study designs make sense when we acknowledge the role study-design plays in the potential for ruling out confounding. One way to see the initial ratings is as bypassing the initial inference one would make in rating a non-randomised study. For example, even if this two stage process was not present, one might initially rate one's confidence in evidence obtained through non-randomised studies as high, on the basis that one assumes the estimate of effect is correct. One would then rate down on the basis of absence of randomisation, before proceeding to rate up on the techniques used to rule out confounding, if present. The GRADE process just bypasses the first stage, by going straight to the middling degrees of confidence warranted by lack of randomisation. On either iteration of the GRADE framework, the end result is the same. One might still call for a small tweak of the GRADE evaluation process that would remove this part of the process, on the basis that it reinforces the idea that there is a hierarchy of study designs. This seems an important point, as such hierarchies do seem to suffer from genuine critiques (e.g., they are less informative than quality assessments; see above and Stegenga (2018, Ch.5)). For this reason only, I would suggest a change to start evidence from both kinds of studies at the same degree of confidence (I develop this idea in more detail in Chapter 9, but it is not necessary to do so to make the point I want to make here). But this is not for the reason that GRADE is less informative when it is a two part design. In either iteration, GRADE is more informative than QATs. It is just that there are benefits to the wider field in not reinforcing the idea that there is a hierarchy of study designs.

5.6 Conclusion

In this chapter I have defended the interpretations of the status and quality evaluative judgements used in EEMM. Specifically, I argued that the status interpretation is justified by the need to hold separate degrees and stability of confidence, and the quality interpretation should be in terms of the stability of confidence. This also makes the pre-2011 GRADE definition of 'quality' correct. To support these claims, I argued that one should hold both degrees and stability of confidence, as they reflect the balance and weight of evidence, respectively. To motivate the EEMM interpretation of quality, I argued that the weight of evidence is contributed to by inadequacies in both the design and implementation of studies, and the analysis of data. Inadequacies pose alternative explanations of the results other than that the hypothesis under consideration is true. Presence of alternative explanations reduces the *net informativeness* of the evidence. As reductions in informativeness typically do not result in evidence that a particular degree of confidence should be selected, what is affected instead is the stability of one's confidence. To link quality to weight, I appealed to accounts of evidence evaluation from philosophy and practice. In both cases, evaluation of quality involves consideration of alternative explanations, and whether net informativeness is reduced. Because of this, quality assessments are evaluating at least one part of the weight of evidence and so quality should be interpreted along pre-2011 GRADE lines. I then argued that as long as quality assessments are seen as the best way to evaluate evidence, then they are no more than weight of evidence assessments. A consequence of this conclusion is that evidence assessments should always be made along the lines of assessing all of what counts towards the weight of evidence. This makes some frameworks that nominally assesses the quality of evidence, but do not take into account all that matters for establishing the weight of evidence, diminished compared with those that do.

Part III

Malleability

Chapter 6

Malleability and medical nihilism

6.1 Introduction

Throughout the chapters that have defended evidence evaluation on EEMM as feasible and conceptually sound, the role of expert judgement in EEMM has been a repeated theme. Expert judgement raises a concern about the influence of subjectivity. A significant recent critique of the methods of contemporary EBM has also raised the spectre of subjectivity. The *medical nihilism* thesis is the radical claim that we should not be confident in the effectiveness of medical interventions, even when we appear to have evidence in favour of effectiveness (Stegenga, 2018). Central to the arguments made in defence of medical nihilism is a concern about the malleability of the methods of contemporary EBM, where to be malleable, a method admits too much room for subjective choices to influence results obtained in implementations of those methods. When subjective choices influence results, those methods can be bent to the needs of interested parties. This is a concern given the immense financial gains that can be accrued through the demonstration of the effectiveness of interventions. The next two chapters will address concerns about the malleability of EEMM. However, EBM+ is the addition of evidence obtained from mechanistic studies to EBM. So, if medical nihilism holds for EBM, then it holds for a significant portion of EBM+ as well. Medical nihilism itself therefore deserves a separate treatment.

In this chapter I reject the medical nihilism thesis. In §6.2, I introduce the *medical nihilism master argument*, and the arguments made for each premise. I then evaluate some criticisms of the thesis in §6.3. I respond on behalf of the medical nihilist to concerns raised with the use of subjective probabilities in the master argument, and with the malleability premise. Importantly, one criticism of the malleability premise is that decisions are actually made more in-line with an EBM+ methodology for causal evaluation. My response to this shows how one cannot just use EBM+ to counter the medical nihilism thesis. However, I find that multiple criticisms of the characterisation of what counts as an effective intervention imply that the medical nihilism thesis does not hold. I then consider whether the malleability premise can support a less strong *medical ambivalence* thesis, which would still pose a problem for EBM. I introduce a novel rejection of the malleability premise in §6.4 based on the claims defended in Chapter 5. Therefore, even a weakened medical ambivalence thesis is false. I conclude that while medical nihilism is false, one can still worry about the malleability of the methods of any methodology for causal evaluation, including EBM+.

6.2 Medical Nihilism

Medical Nihilism is the thesis that we should hold low confidence in the effectiveness of medical interventions (Stegenga, 2018). This is even the case when evidence appears to point strongly in favour of the intervention. The argument for this thesis relies on the use of Bayes theorem:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{6.1}$$

The Bayesian tradition covers a variety of fields and Bayes theorem, a consequence of the axioms of probability theory, can be used for a variety of means. The next chapter makes extensive use of *Bayesian epistemology*, which holds that rational degrees of belief are probabilities, and seeks to use this interpretation of belief to answer the question of how strongly one should believe in a proposition. Many but not all theories of Bayesian epistemology utilise some form of Bayes theorem. What Stegenga calls the *master argument* for Medical Nihilism uses Bayes theorem as a means to make an inference about the confidence one should hold in the effectiveness of interventions. It does so by making arguments for assignments of values to the probabilities in the three terms on the right hand side of eq. (6.1). This use of Bayes theorem is squarely in the field of *Bayesian confirmation theory*, which offers a theory of scientific inference based on the theorem. However, only a basic understanding of the terms in the theorem is needed for this chapter, so I will briefly introduce them in the remainder of this section. While introducing the terms I will also summarise the reasons he gives for assigning probabilities to terms in his master argument.

In Bayes theorem, $P(E|H)$ is called the *likelihood*, and stands for the probability of obtaining evidence E for a hypothesis H given the truth of H. In other words, how probable the evidence is on the assumption that the hypothesis is true. For example, for H='drug D reduces blood pressure', if it were true that D reduces blood pressure, then with high probability we would expect to observe E='a decrease in blood pressure' in patients that take D, compared with those that don't take D: $P(E|H)$ would in this case be high. Stegenga argues that $P(E|H)$ should be low for all new interventions. He argues that where evidence does demonstrate effectiveness of an intervention it is typically evidence of *low effectiveness*. His evidence for this is the ubiquity of small effect sizes. He reasons that if one's hypothesis is that an intervention is effective, then the likelihood should only be high if one is likely to observe a large effect size. But as we seem to only ever observe low effectiveness, $P(E|H)$ should be low.

$P(E)$ stands for the *prior probability of the evidence*, and assignments of probability reflect one's expectation of obtaining that evidence, independent of the truth of the hypothesis. Consequently, high values for $P(E)$ denote evidence that one would expect to obtain, independent of whether any hypothesis is true or not. For example, one would expect to observe a difference in blood pressure between intervention and control group if the treatment group of the trial was composed entirely of people with low blood pressure, and the control group composed entirely of people with high blood pressure. Independent of whether H is true, one would expect to obtain E in this set up. This illustrates why high values for $P(E)$ are typically associated with 'bad' evidence. Stegenga argues that because of the need to make many subjective choices when carrying out clinical trials $P(E)$ should be high. Trials are at risk of many biases, presence of which may mean the results of such trials are not representative of any true causal relations. Moreover, biases will often over-estimate observed effects. Compounding this issue is that experimenters must make many choices when designing and implementing trials, and those choices can introduce biases. Stegenga argues this makes the methods of EBM *malleable*, which is a significant problem given the great financial interest involved in whether a treatment is effective. There is motivation to commit fraud and the malleability of methods means that an interested party could design a trial to obtain favourable results. So we should expect to observe greater benefit in the intervention group of a trial than in the control group, regardless of whether interventions are truly effective. Therefore, $P(E)$ should be high.

$P(H)$ is called the *prior probability of the hypothesis.* This stands for the degree of belief one holds in the truth of H before one obtains some evidence E. $P(H)$ is obtained conditional on all of one's background information B about the hypothesis, so it is really $P(H|B)$. To be precise, all of the terms are conditional on B, e.g., $P(E|H, B)$, $P(E|B)$. For simplicity, the B is dropped. E is then some new evidence that one is entertaining as support for H. For example, when investigating blood pressure drug D, one might have a wealth of pre-clinical evidence that indicates D will be effective. It may be of a similar class to drugs already proven effective, or D may have been tested in other clinical trials (e.g., Phase I and II trials that are used to test safety and efficacy but can't alone ground effectiveness claims). One would therefore hold a high $P(H)$. The converse also holds. If one had no evidence about the effectiveness of D before running the trial, $P(H)$ would be low. One then runs an experiment to obtain new E. Evidence for $P(H)$ would also include more general information about whether interventions supposedly proven effective are subsequently found to be ineffective. If this phenomenon is common, then one would start to doubt whether any intervention proven effective would remain so in future. Stegenga claims that there are many instances of this phenomenon. So he reasons that it is unlikely that any new intervention is effective, even if there is pre-clinical evidence suggesting it is. Therefore, $P(H)$ should be low.

Putting the terms together, $P(H|E)$ should be low. This means that even when E is in favour of H, you should not believe the intervention is effective. This is a consequence of the Bayesian calculation. If 'low' means a probability less than or equal to 0.2, and high means a probability greater than or equal to 0.8, plugging these values into eq. (6.1) gives:

$$P(H|E) \leq 0.05 = \frac{0.2 \text{ x } 0.2}{0.8}$$

So, the probability of the hypothesis conditional on the evidence should always be low. If this thesis is true it radically redefines the way we think about medicine. It would mean that all research into interventions is in vain, as any evidence we do obtain will be uninformative about whether interventions are in fact effective. For the specific concerns of this thesis, it is important to consider medical nihilism, as it may make the task of evaluating causality fruitless.

6.3 Criticisms

In this section I consider a number of criticisms of medical nihilism. I argue that in the face of these critiques, the medical nihilist can sustain both the use of subjective probabilities in the master argument, and claims about the malleability of EBM. However, they will struggle to respond to critiques of: i) what evidence counts towards setting $P(H)$; ii) what 'effectiveness' means in H. The consequence of these critiques is that $P(E|H)$ and $P(H)$ should not always be low, rendering the claim that $P(H|E)$ is always low, and so too medical nihilism, false. However, because the malleability premise can be sustained, there is a weaker, *medical ambivalence* thesis that can be argued for. Whether this weakened thesis holds depends on the malleability premise so I move on to a rejection of that premise in the next section.

The first kind of criticism takes aim at the nature of the probabilities employed in the master argument. Gillies (2019b) argues that the use of Bayes theorem in the master argument is inappropriate as the probabilities utilised are purely subjective. The values subjective probabilities take are entirely up to the person eliciting them, in this case Stegenega. Gillies charges this with arbitrariness as they can take on wildly different values from person to person. He contrasts them with objective probabilities, typically derived from observed frequencies (see Hájek (2008) for a run down on the interpretations of probability). A calculation made on objective probabilities is deemed reliable by Gillies, as the values of those probabilities should not vary depending on the person running the calculation. As Stegenga's calculation of the posterior probability is based on his own subjective degrees of belief for each of the terms in the equation, his conclusion is arbitrary.

One can doubt whether the subjectivity of Bayesian inference is that much of a problem. For example, Sprenger (2018) argues that on some senses of objectivity, subjective Bayesian inference is no more subjective than Frequentist inference, which makes use of objective probabilities. He also argues that subjective Bayesianism can help to ensure a kind of objectivity, namely *interactive objectivity*, where the objectivity of science is secured by the ability to communicate and critique judgements. Subjective Bayesianism fulfills this sense of objectivity as it allows assumptions to be made clear at all stages of inference. So one might doubt that the master argument loses all force on account of its use of subjective Bayesian inference. To level interactive objectivity in support of the use of subjective probabilities in his master argument, Stegenga would need the reasons for his assignments of probability to terms in the master argument to be critiqued and justified. There are however two potential problems with the arguments he makes to do so. One is a problem concerning the *malleability premise*, and the other is with the characterisation of H.

The malleability premise is used to justify assigning high values to $P(E)$. One criticism of it highlights the restrictive nature of Stegenga's characterisation of contemporary medicine. When he argues that the methods of contemporary medicine are too malleable, he is in fact arguing that the methods of *contemporary EBM* are too malleable. This is because he focuses on how those methods are prone to bias. One objection to this is that there is an alternative methodology for causal evaluation, namely EBM+, that he does not consider (Williamson, 2020). Moreover, it seems to be the case that decisions in medicine are made, *at least implicitly*, on the basis of multiple lines of evidence (Devanesan, 2019; Williamson, 2020). This is because the production of guidelines on interventions explicitly follows the EBM approach to evidence assessment, but in practice, judgements will be informed by a number of different lines of evidence, including the plausibility of mechanisms. So the overall evaluation of an intervention's effectiveness includes evidence from mechanistic studies. The decision making process thus seems to lean more towards the EBM+ approach.

This is significant, as Williamson (2020) defends EBM+ against a charge of malleability. So evidence evaluation actually proceeds in a fashion more akin to a methodology that is plausibly not at risk of malleability. He evaluates a number of responses to the charge of malleability, and finds in favour of one based on evidential diversity. The basic idea is that mechanistic studies and clinical studies are each at risk of different kinds of biases. If one is concerned with bias in clinical studies, obtaining more evidence from the same kind of methods is not likely to remove this worry. Obtaining evidence from mechanistic studies that points in the same direction as the evidence from clinical studies should lessen worries about the risk of either method being biased. This is because the strengths of each kind of evidence can help to mitigate the effects of the other kind's weaknesses. These weaknesses will include biases, so malleability is "not a substantial problem for EBM+" (Williamson, 2020, p.13).

One worry about this argument relies on a distinction between *methods*, *meta-methods*, and *methodologies for causal evaluation*. EBM and EBM+ are competing methodologies for causal evaluation. The methods of EBM and EBM+ are the various kinds of clinical and mechanistic studies, respectively. Meta-methods are used to assess a body of evidence composed of individual studies, each of which will have implemented a particular kind of method. What is assessed is whether claims made on the basis of those studies are justified. Some meta-methods will assess whether the studies were designed, implemented, and their results analysed, in accordance with specifiable standards, e.g., quality assessment tools (QATs), GRADE. Other meta-methods may assess the results of studies in an evidence base to derive more precise conclusions, e.g., meta-analysis. These examples of meta-methods are firmly in the EBM tradition. As already noted in this thesis, there are few examples of evaluative frameworks in the EBM+ tradition, but one EBM+ meta-method is EEMM. It is important to acknowledge this distinction, as Stegenga's arguments only really work against meta-methods. It is certainly the case that all methods are at a risk of bias, but it is not the case that all implementations of such methods result in biased studies (Devanesan, 2019). Clinical trials are at a risk of all kinds of bias, and it is the task of the design and implementation of individual trials to control for it. So the potential for bias is not an argument for the malleability of medical methods, contrary to what Stegenga claims. Devanesan (2019) argues that what matters is context. Whether trials are biased should be determined on a case-by-case basis. Stegenga anticipates this move and also includes critiques of the meta-methods of EBM, namely evidence hierarchies, meta-analyses, and quality assessment tools (QATs). If these meta-methods are also malleable, then whether the studies that provide evidence are biased is hard or impossible to tell. The claim that the meta-methods of EBM are (too) malleable is where the force of the argument for malleability comes from.

The problem with Williamson's argument above is that it is about the malleability of EBM+ as a *methodology for causal evaluation*, rather than any *meta-method for evidence aggregation or evaluation* built on EBM+ principles. The response to the worry about the malleability of EBM+ is that the prescription to include both kinds of evidence in an evaluation of causality will decrease the risk of malleability. It does not say anything about the malleability of the meta-methods of EBM+. So one might still have malleability worries when it comes to how an EBM+ approach to evaluating causality would use meta-methods to evaluate evidence. This is not just a concern about the specific EEMM process. It is informed by more general considerations about the nature of evidence evaluation performed along EBM+ lines. As pointed out in Chapter 4, the methods employed in mechanistic studies are highly heterogenous. This introduces the potential for many kinds of biases, and for evaluation of evidence obtained from mechanistic studies to require many subjective choices. I raised this concern about expert judgement in the same chapter, and judged it to be potentially ineliminable. But this may lead us to conclude that the meta-methods of EBM+ are too malleable, and at risk of influence by interested parties. I defer responses to this worry about the malleability of EEMM to later chapters, but it suffices here to note that the malleability premise of medical nihilism cannot be rejected just because evidence evaluations proceed along lines more akin to EBM+.

It may however turn out that EEMM is not at risk of malleability, and I will argue for ways to reduce such risk in Chapters 7 and 8. The medical nihilist might however argue that even if this is so, the fact that decisions are made *implicitly* in line with EBM+ would make such decisions even more prone to influence from interested parties. Indeed, as Plutynski (2018, p.130) points out, in the case of exposure assessment (see §3.1), industry can obfuscate and delay decisions by making demands for more mechanistic evidence when there is evidence of a strong association between a chemical and a harm. This could also be the case in intervention assessment, but with industry instead levelling mechanistic evidence in support of the effectiveness of the interventions they are interested in. They will likely have some pre-clinical data, including evidence that suggests a mechanism of action for the intervention. The problem with this is that if this evidence is not explicitly evaluated, then it is only utilised to support decisions in an opaque way. Charges of subjectivity and influence can then easily be levelled at such a process. So a rejection of the malleability premise on the grounds that it is the methods of contemporary EBM that are criticised, not EBM+, does not work.

Two final problems with the medical nihilism thesis derive from issues with how the hypothesis in the master argument is characterised (Broadbent, 2019; Devanesan, 2019; Gillies, 2019b). Recall that Stegenga's hypothesis is about any new intervention, but the evidence that counts for setting $P(H)$ low refers to all medical interventions. However, there are many counterexamples of effective interventions, a number of which Stegenga acknowledges. So he cannot claim that almost all past medical interventions are ineffective. That there are counter-examples that he does not acknowledge implies that there may be even more. This means the assignments of values to $P(H)$ should not always be low, but should depend on the background knowledge that supports H.

To counter this issue, the medical nihilist could attempt to find a way to mark out which class of interventions we should be considering. Another option is to undertake a more detailed analysis of the proportion of ineffective to effective treatments that shows why $P(H)$ should be low. Even if this could be achieved there is another problem: it is not clear what is meant by 'effective' in the hypothesis. As Broadbent (2019) argues, Stegenga's notion of effectiveness is actually a notion of 'high effectiveness'. Stegenga discounts small effect sizes, and claims that there are few or no 'magic bullets' in contemporary medicine, where a magic bullet is an intervention specifically effective for the biological basis of a disease, that also comes with few side effects. But there is more to effectiveness than this. One might not cure some disease, but a treatment can still be effective. Equally so, the presence of side effects, even severe ones, may not outweigh the benefits of an intervention.

Both these features are highlighted by Devanesan (2019) in an example from contemporary medicine that Stegenga ignores, namely, the use of anti-retrovirals to prevent death from HIV-AIDS. Now, anti-retrovirals do not cure the patient of HIV, as the virus remains in the body. It is also the case that there are significant side-effects from taking the anti-retrovirals. Stegenga might therefore consider it ineffective, or at the very least not a magic bullet. But they do prevent progression to AIDS and an untimely death, and have played a part in stopping the AIDS epidemic in most high-income countries. If this is not a case of an effective intervention then it seems like the game is rigged by the medical nihilist. This problem with how 'effectiveness' is characterised would mean that $P(E|H)$ is not as low as suggested by Stegenga. When the hypothesis is about a less restricted notion of effectiveness, then it should be expected that small effect sizes will be obtained.

While the malleability premise may still hold, the problems facing the characterisation of the hypothesis suggest medical nihilism is false. The values for both $P(H)$ and $P(E|H)$ can be contested, and a case made that they should be higher. However, if it is true that because of malleability worries $P(E)$ is always high, then confidence in the effectiveness of medical interventions will only be moderate-to-high when both $P(H)$ and $P(E|H)$ are also high. Such a situation seems unlikely to hold in general as $P(H)$ would typically be non-committal for novel interventions (e.g., in the range [0.4,0.6]). This is because there would be pre-clinical data showing potential effectiveness but little direct evidence about that intervention. In most cases, where $P(H)$ is non-committal, $P(H|E)$ would be low-to-moderate. So we might not be nihilists, but we should be at best *ambivalent* about most new drugs, except when $P(H)$ is high. This *medical ambivalence* thesis would also pose a problem for medicine as it implies that we can only be confident that a drug works when we have a large amount of data about the intervention (because of requiring high $P(H)$). Moreover, on this view, there is no general stipulation that we should expect all of our interventions to be ineffective. So a consequence of medical ambivalence is that it remains possible that we end up employing some interventions in general practice a reasonably large amount of time after we could have first recommended them. The methods of EBM could have demonstrated effectiveness, but we delayed such decisions because of concerns about the *general* malleability of those methods, which kept $P(E)$ high. To dismiss medical ambivalence I therefore look at a way to reject the malleability premise.

6.4 An argument against the malleability premise

In this section I reject the malleability premise on the grounds that Stegenga's argument that the methods of EBM are too malleable is not made against the *best* methods of EBM. In particular, evidence hierarchies are not used to evaluate evidence in practice, and QATs do not evaluate all that is relevant to the quality of evidence.

Recall from the previous section that Stegenga actually directs his arguments about the malleability of EBM at its *meta-methods*, namely evidence hierarchies, QATs, and meta-analyses. He is right about hierarchies: they are an inadequate way to assess evidence. However, the kind of hierarchy he considers is a crude one, in that the ranking of evidence is performed by consideration of only one property, namely the ability of a method to rule out confounding. His argument is that ranking by one property is less informative than an assessment by many properties (this argument was introduced in §5.5). But it is not clear that major organisations use only a simple hierarchy to evaluate evidence. For example, the National Institute for Clinical Excellence (NICE) use quality assessment tools, which do assess by many properties: when producing guidelines to assess the quality of evidence, they use a combination of the Cochrane Collaboration systematic review guidance and the GRADE framework (National Institute for Clinical Excellence, 2012; Thornton et al., 2013). On the other hand, the European Medical Agency (EMA) do not use a specific quality tool, instead opting to evaluate clinical efficacy by considering only the results of RCTs. However, evaluations are more detailed than just what kind of clinical trial method was used. Sample size, method of randomisation, sub-group analyses, and outcome appropriateness are all considered when evaluating efficacy (see, e.g., European

Medical Agency (2018, pp.58-59)). Moreover, the kind of evidence hierarchy Stegenga rails against is typically used only for pedagogical purposes, or when communicating the very basic idea behind EBM's theory of evidence. In reality, the evaluation process is explicitly much more nuanced than just whether a particular kind of clinical trial method was used to obtain the evidence.

Stegenga does however turn his attentions to quality assessment in the form of QATs (introduced in §5.5). Although he thinks they are better than evidence hierarchies, he concludes that they are also at too great a risk of malleability. However, as I identified in §5.5, the QATs he focuses on are really only risk of bias tools. And I also argued in that chapter that an evidence assessment that looks at the weight of evidence is more informative than one that looks solely at risk of bias. GRADE is an example of a framework that evaluates more of what is relevant to the quality of evidence, but he does not analyse whether GRADE is at risk of malleability. So on Stegenga's own terms, the QATs he thinks are unreliable are not the best way to assess evidence. The claim that the methods of EBM are too malleable depends on identifying malleability in the *best* methods of EBM. Stegenga's argument does not do this, making the claim unwarranted.

One might worry that GRADE is also too malleable. Indeed, there are two concerns Stegenga has about the malleability of QATs that may transfer over to GRADE. One is that of *inter-rater* and *inter-tool* reliability, on both of which QATs perform poorly. When an evaluation framework is applied by more than one evaluator to the same evidence base, and the evaluators rate the evidence similarly, then the framework has good inter-rater reliability. When one evaluator applies more than one framework to the same evidence base, and the outcome is similar, then the frameworks have good inter-tool reliability. Low reliability on both counts implies that evaluators are a major factor in the outcomes of assessments. There is some evidence that GRADE does well on inter-rater reliability. Mustafa et al. (2013) found good inter-rater reliability amongst experts given the same systematic review to apply the GRADE process to. Moreover, reliability was higher when applying the GRADE process than when the experts were asked to use their own background knowledge to assess the evidence. Kumar et al. (2016) demonstrated similar results. Additionally, Mustafa et al. (2013) found that for a group of students untrained on GRADE, inter-rater reliability improved 6-fold after undertaking a short training programme on the GRADE framework. This suggests that problems with inter-rater reliability can be (at least) partially mitigated by thorough training on the framework.

There is however little evidence on inter-tool reliability. Such evidence may be difficult to obtain given that existing frameworks may differ substantially from GRADE. Indeed, GRADE improves on many existing frameworks by evaluating more that is relevant to the quality of evidence. This was the case with QATs, but will also be the case for other frameworks that focus on the risk of bias. Because of this, one might actually expect outcomes of GRADE evaluations to differ relative to existing QATs. One reason to think this is that GRADE includes explicit evaluation of ‘publication bias’, which involves consideration of the funding source of trials. This is not something that many QATs will look at. So GRADE would likely rate a body of evidence at risk of publication bias lower, all else being equal, than a QAT would. So it is hard to say whether inter-tool reliability would even be a good method for evaluating GRADE as a quality assessor. Additionally, the presence of biases in trials was not enough to consider malleability a concern for Stegenga, as there also had to be the potential for influence by interested parties. GRADE explicitly heads off this concern by the inclusion of publication bias. So even if one is still concerned about inter-tool reliability, there is reason to think GRADE can head off the fraud aspect of malleability concerns.

A final refuge for the proponent of the malleability premise is what Stegenga calls the under-determination of evidential significance (UES). He levels this to counter a potential objection to his claims about the inadequacies of QATs, namely that there may be, in future, better ways to assess evidence than QATs. He counters this objection by arguing that quality assessments will always be inadequate because there can be no objective means by which to weight the impact of quality criteria on judgements of the strength of evidence. So even if there is a better quality assessment framework out there than QATs, and I have argued that there is, this will not matter for the general point he makes about the malleability of quality assessment. This however seems like a strange point to argue. As UES is a general principle about evidence in science, it seems like this argument would extend to other fields. Why should UES just hold for medicine, rather than, say, physics? If it is the case that this principle holds in general then we might be tempted to be physical nihilists, or chemical nihilists. But to say that we should hold low confidence in our best physical theories seems too strong a case to make. Now one might think medicine is a less complete a science than physics. So at present it might be the case that our best methods are not nearly as good as those of physics. But this does not alone rule out the possibility of improving medical methods, whereas UES does. So Stegenga cannot level UES to support a general attack on the (meta-)methods of medicine.

To sum up, evidence hierarchies are not serious contenders to be classed as meta-methods, and QATs are less informative than other assessment frameworks. Stegenga is therefore not arguing against the best meta-methods of EBM, so any claims of their malleability fall flat. There is some truth to his claims about meta-analyses. The implementation of their methodology does require subjective choice. But the consequence of this claim has been challenged by Holman (2018), who argues that the successive application of meta-analyses to the same question cancels out the problems introduced by subjective choices. There is of course no response to the charge that subjective choices are required in the implementation of all the (meta-)methods analysed. This is something the GRADE working group admit and welcome, but as Sprenger (2018) argues, the mere presence of subjective choice does not automatically lead to subjective inferences. In other words, malleability should still be taken seriously as a worry for each method and meta-method, but there is not enough reason to think that *in general* medical methods are so malleable that $P(E)$ should be high for all new interventions. So even a medical ambivalence thesis is false.

6.5 Conclusion

In this chapter, I considered criticisms of medical nihilism and found that a criticism of the hypothesis employed in the Bayesian master argument holds while criticisms of both the use of subjective probabilities and the malleability premise do not. Because of this, I found medical nihilism false, but considered a less strong medical ambivalence thesis to remain plausible. This thesis ultimately depends on the malleability premise being true. I argued that it is not true, as the meta-methods that Stegenga argues against are not the strongest meta-methods available. Moreover, there are reasons to think malleability is less of a concern on the best meta-methods. However, I did concede that malleability may be a worry context-to-context. Instead of a wholesale rejection of the effectiveness of medical interventions on the basis of malleability concerns, what matters is evaluating the risk of malleability context-to-context. This means that it is applications of the methods and meta-methods of a methodology for causal evaluation in medicine that are at risk not the methodology as whole. While EBM+ as a methodology for causal evaluation may not be at risk of malleability, its meta-methods may be. I therefore move on to developing a way to head off malleability concerns about EEMM in the next chapter.

Chapter 7

Malleability and formalisation

7.1 Introduction

In Chapters 3 and 4 I identified that expert judgement is required when carrying out an evaluation on EEMM. In particular, evaluators must make many choices when deciding how evaluative criteria affect evaluative judgements. This raises a worry about malleability. Subjectivity will have a malign influence on judgements when one can choose what effect evidence has on one's confidence presumably on a whim. It would then be possible to actively bias the evaluation. I also raised the issue of the potential malleability of the meta-methods of EBM+ in the previous chapter. There I rejected Williamson (2020)'s claim that being evidentially diverse ameliorates malleability concerns about EBM+. Another route to ameliorating malleability that Williamson (2020) considers, but ultimately rejects, is the prospect of formalisation of the evaluation process. This may however be more relevant to meta-methods than it is to EBM+ as a methodology for causal evaluation. Currently, EEMM is mainly qualitative, asking users to evaluate their confidence without use of formal models. Applying formal tools to an evaluation process offers a means to restrict the influence of subjective choice on inferences. For example, he considers the *E-synthesis* model of evidence aggregation, which uses a Bayesian network model to formalise Hill's criteria for causality (Landes et al., 2018). This model elicits subjective probabilities from experts about the relevance and reliability of evidence reports. These probabilities then feed into multiple *indicators of causality*, which are modelled on Hill's criteria for causality, and include both correlations and mechanisms. The model is then used to calculate a posterior probability for a causal claim. The use of the formal model is supposed to make clear what judgements about evidence are made, and how they should influence final judgements about causality. Evaluators cannot make inferences to conclusions, from relevance and reliability reports, that violate the norms of the model. If one accepts the norms, then those inferences are deemed reliable. However, as this example shows, there will be a need for subjective choices when both eliciting probabilities, and when making model assumptions. Formal models are thus open to a charge of subjectivity and Williamson rejects them as a means to reduce the malleability of EBM+.

I have contended that the mere use of subjective probabilities is not a problem. What matters instead is whether one can make the judgements for assignments of probabilities transparent, and therefore contestable. It is also the case that Williamson restricts his analysis of formalisation to concrete examples of formal models that are appropriate for formalising the EBM+ approach. He does not analyse whether the EEMM approach can in principle be formalised and whether this would benefit the process. Moreover, the use of subjective Bayesian inference is not the only way to formalise the process. One avenue to explore is competing models of Bayesian inference that apply normative constraints on permissible degrees of belief. One motivation for such models is that the additional constraints make degrees of belief more or less objective than one another. They thus hold promise for the ability to rule out subjective influences, in virtue of their core goal being to limit subjectivity in Bayesian inference. For both these reasons, formalisation remains a live option.

In this chapter I argue that formal models of belief can be a means to reduce the influence of subjectivity on an evidence evaluation by the EEMM framework. In particular, I argue that one theory of Bayesian epistemology, namely objective Bayesian epistemology, is the right way to formalise the inferences made on EEMM. In §7.2, I introduce a number of theories of Bayesian epistemology. I delineate these varieties along two different dimensions. Along one dimension are the norms that govern how evidence imposes constraints on belief: different theories impose norms that make permissible degrees of belief more logically objective (§7.2.2). The motivation for delineating along this dimension is that such norms hold promise for restricting the influence of subjective choice in the part of the EEMM process that uses evidence to impose constraints on belief. Another dimension along which theories differ is on the precision with which graded beliefs are represented (§7.2.3). I outline the benefits and drawbacks of each theory, but do not make a judgement on which one is *in general* the correct Bayesian epistemology. Instead, in §7.3, I use these benefits and drawbacks to determine the theory best suited to be applied to EEMM. This decision is made in a pragmatic manner, by matching the needs of the application to the benefits and drawbacks of each model of belief. Finally, I consider an objection to formalisation: that formal models are hard to implement, which results in needing to make many idealisations, and many choices (§7.3.2). This may make judgements arbitrary and open to the influence of subjectivity. I argue that formal models result in less subjectively influenced outcomes than non-formal models, but that application should involve approximations to idealised models. Moreover, we should recognise, rather than ignore, the extent of subjective choice needed to carry out each evaluation. This will require a parallel evaluation of subjectivity.

7.2 Evidential constraints on formal models of belief

In this section I introduce details of, and motivations for, varieties of Bayesian epistemology. In the task of applying formal models to EEMM I take a pragmatic stance, where what model is right for EEMM depends on the practical needs of the evaluation process. Each model will have its benefits and drawbacks that one might argue makes it the normatively correct theory of Bayesian epistemology. But in a particular application it can be the case that the relative importance of certain benefits or drawbacks will change. A formal model may, for example, do well in terms of accuracy but be criticised on grounds of practical expediency. This kind of general criticism is however less of an issue if a practical application of this model values accuracy over expediency. This may be especially so if the risks associated with being inaccurate are severe. In the following I will thus make clear what the benefits and drawbacks of each model are, relative to benefits and drawbacks of rivals, without making a general argument in favour of any one model.

7.2.1 Bayesianism

Bayesian epistemology holds that beliefs come in degrees, and the strengths of those beliefs should satisfy the laws of probability. Within these constraints, one holds *prior* degrees of belief in propositions, before learning something that *updates* those prior beliefs to *posterior* beliefs. This is the basic apparatus of all of the many kinds of Bayesian epistemology. Different theories of Bayesian epistemology apply and supplement the basic apparatus in a number of ways to answer the question of how strongly one should believe in the propositions one entertains. One important way theories can differ is on how a theory adds to the basic apparatus, such that permissible prior degrees of belief can be further constrained. Another way is on the precision with which degrees of belief are represented by probabilities. These distinctions cut across one another and I will expand on each in turn.

7.2.2 Subjectivity and Bayesian epistemology

A major criticism of the Bayesian approach is the *problem of the priors*, which charges Bayesian inference with being too subjective. A version of this criticism was also made against the master argument for medical nihilism by Gillies (2019b) (§6.3). The general argument is that inference in most applications, and particularly within the sciences, needs to be objective. And Bayesian inference is too sensitive to prior degrees of belief, which according to the core norms of Bayesianism need only satisfy the laws of probability. Within that one constraint, scientists can choose whatever degree of belief they want for their priors. Updating on subjective degrees of belief results in subjective posterior beliefs. So Bayesian inference is too sensitive to the subjective choices made when one assigns values to prior degrees of belief.

In response to this, different theories of Bayesianism add to the core norms. These additional norms supply constraints on permissible degrees of belief with the intention of ensuring the objectivity of Bayesian inference. However, varieties that mean to ensure the objectivity of Bayesian inference are still split into camps of subjective and objective Bayesians. This is because of variable usages of the terms subjective and objective. On the one hand, talk of the objectivity of science can mean a number of things. It can mean that there is a way the physical world is and our theorising must accurately represent it. It may mean that our inferences can only depend on physical evidence, and not mental entities like degrees of belief. It may also mean inter-subjective agreement. There are also many other senses of objectivity in science (see Sprenger (2018) for an overview of them in the context of subjective Bayesianism). On the other hand, the sense of subjective and objective used to describe varieties of Bayesianism has to do with *logical* objectivity. A method for assigning degrees of belief is logically subjective when it allows for at least two agents to hold different degrees of belief, given the same body of evidence. A logically objective method for assigning degrees of belief will allow only one correct degree of belief to be held by any number of agents, given the same body of evidence. In the following discussion when I talk of objectivity, I talk of it in the sense of logical objectivity.

Because of this, when I talk about whether permissible degrees of belief on some theory of Bayesian epistemology are more or less objective, I am not talking about whether those degrees of belief accurately represent the 'objective' state of the world. One might wonder then whether use of such norms can head off concerns about malleability. Conclusions reached by application of, say, an 'objective Bayesian' method (introduced in more detail below), may only reflect the norms of the method rather than the causal relations that actually hold. However, malleability is a concern about the influence of subjective choice, not a concern about the accuracy of a method. The purpose of securing logical objectivity is to reduce the influence of subjective choice on assignments of values to prior degrees of belief. Whether a method is accurate is another consideration to whether it is malleable. It is entirely possible that an accurate method may be at risk of being malleable, while an inaccurate method may admit no subjective choice. So securing logical objectivity, by means of removing the influence of subjective choice, is a way to ameliorate malleability, even if it does not secure accurate judgements. It is not the case either that this means an objective Bayesian account is automatically the set of norms to apply to EEMM. There are nuances to subjective and objective Bayesian accounts that must be accounted for before making such a judgement.

Each variety of subjective Bayesianism will hold that there is at least one constraint on permissible degrees of belief, namely probability theory. Varieties are then distinguished by what additional constraints are allowed. In a survey of the development of subjective Bayesianism, Joyce (2011b) identifies two kinds of subjective Bayesian, the subjective and the *tempered*. The subjective Bayesian is allowed to select any degree of belief from those that are consistent with probability theory *and* one's evidence. The tempered Bayesian also requires no constraints on prior degrees of belief. Objectivity is ensured instead by appealing to convergence theorems: as more and more data is accumulated, the influence of subjective priors on posteriors is 'washed out', as the different priors held by different agents converge on the correct posterior belief (see Howson (2003); Howson and Urbach (2006)). Convergence on the correct posterior belief only happens asymptotically, so in most practical situations there will not be one correct probability function picked out. In the present context then, tempered Bayesianism can be lumped in with subjective Bayesianism, as they both require no additional constraints beyond consistency with probability theory. Williamson (2010) opts for two varieties of subjective Bayesianism: *strictly subjective Bayesianism*; and *empirically based subjective Bayesianism.* The former is only constrained by the laws of probability theory while the latter *calibrates* degrees of belief to evidence, which makes it similar to Joyce's subjective Bayesian. I expand on this idea below. Combining these kinds, we can identify three broad varieties: the strictly subjective, the tempered subjective, and what I call the *calibrated subjective Bayesian.* This final variety of subjective Bayesian will use empirical evidence to assign degrees of belief.

Adding a norm that requires calibrating to empirical evidence is intended to ensure the objectivity of Bayesian inference, as it requires calibrating to what are typically called 'objective probabilities'. Following Williamson, I will call them physical probabilities, so as not to confuse them with the sense of logical objectivity used in this chapter. Still, the values these physical probabilities take are considered independent of the subjective choices of agents. It is important to distinguish here between *interpretations of probability* and *theories of Bayesian epistemology.* There are three kinds of physical interpretations of probability, and each hold that probability is a physical quantity. There are limiting relative frequency theories (e.g., Von Mises (1982)), propensity theories (e.g., Gillies (2000)), and physical chance theories (e.g., Lewis (1980)). Physical chance theories are single case, in that physical probabilities attach to *an* event. Generic probabilities by contrast are attributes of repeated events. Limiting relative frequency theories are generic, while propensity theories can be both generic or single case. We might want to say that there is a physical chance for each proposition we entertain, and this is what we are using evidence to calibrate our degrees of belief to. For example, there is a mind-independent physical chance that a flipped coin will land heads. For a fair coin the chance is arguably 0.5. This idea was introduced in §5.4.1 as the *objective chance* of an hypothesis. I use physical chance for similar reasons as I do for using 'physical probability' instead of 'objective probability'. There are two questions to answer here. One, how do we gain evidence of physical chance to calibrate to? Two, if we calibrate to physical probabilities, would we remain a subjective Bayesian?

One way of constraining degrees of belief through evidence of physical chances is the *Principal Principle*, which states that when your evidence for p is that the physical chance Ch of p equals x, then your degree of belief b in p should equal x: $b(p) = Ch(p) = x$. We can gain access to physical chances through frequency data. To do so requires finding the limiting relative frequency of an event E and identifying it with the physical chance of E, where the limiting relative frequency of E is the proportion observed of E out of the total number n of events observed, as the size of n approaches infinity. *Frequency data* is composed of frequencies obtained in a finite, rather than asymptotic reference class. Justifications for both identifying chance with, and for inferring from finite data to, limiting relative frequencies, are based on the law of large numbers. Such justifications go beyond the scope of this chapter, but details can be found, e.g., in Howson and Urbach (2006, Ch.3f) and (Mayo, 1996, p.162-173). Another way evidence can constrain belief is Williamson (2010)'s *Calibration* principle, which again identifies those probability functions compatible with one's evidence. It does however allow for more complex evidence of chances, through constraining belief to both physical data and to a set of 'structural constraints'. These constraints allow, for example, causal information to constrain what counts as a permissible degree of belief. This is a brief summary of a more complex principle (for details see Williamson (2010, pp.39-49)). Whichever principle or norm one chooses, the point is that a calibrated Bayesianism ensures that permissible degrees of belief in an operating context are constrained by the evidence of physical chance available.

One might think that this is enough to ensure objectivity. Why is it then that one can calibrate to physical probabilities and remain a subjective Bayesian? One might think that as it is evidence of physical chance that constrains permissible degrees of belief, subjective choice does not enter the equation. However, in most cases, a calibration norm will only select a set of probability functions. Often, our evidence of physical chance is only approximate, e.g., when frequency data is only consistent with an interval of probabilities. And without further norms, the degree of belief one selects from within this set will depend upon subjective choice. An objective Bayesian would say that one needs further norms to select degrees of belief from those compatible with the evidence. Moreover, these norms would be *non-evidential* in nature.

In Williamson (2010)'s objective Bayesianism, to select a degree of belief from those compatible with both the laws of probability and one's evidence requires applying an *Equivocation* norm. This norm requires degrees of belief to be *sufficiently equivocal*: when evidence fails to determine degrees of belief, one should select a belief function that is sufficiently close to an equivocator function. The equivocator function is determined by giving each basic possibility one is entertaining equal probability. Closeness to the equivocator is measured by Kullback-Leibler divergence (for formal details see Williamson (2010, pp.28-29)). This process may not pick out a single sufficiently equivocal function in all cases. Selecting a degree of belief from within the set of sufficiently equivocal functions is then to be determined pragmatically. Williamson (2010, p.163) notes that what it takes to pragmatically determine this function is a matter for further investigation. But objective Bayesianism still has some benefits over subjective calibrated Bayesianism. When evidence does not determine degrees of belief, one applies the equivocation norm, rather than selecting a degree of belief through one's expertise. This limits the risk of subjectivity. Moreover, this norm is independently motivated. Williamson argues it is the most cautious policy to adopt when making decisions. This leads to less risky decisions on average than if one is a (calibrated) subjective Bayesian.

7.2.3 Precision of degrees of belief

Objective Bayesianism can however be charged with going beyond the evidence. A critic may argue that selecting the most equivocal function introduces too much information: often the information we do have is incomplete and approximate and choosing a point value goes beyond this information (Kyburg, 2003). The choice of an equivocation norm, or the selection of the function that is sufficiently equivocal may also come down to a matter of subjective choice. The best one may be able to do in the face of incomplete and approximate evidence is to hold similarly unspecific beliefs. Indeed, this is a prime motivation of *imprecise probabilism.* A version of this normative theory of belief holds that one should interpret one's belief state as a set of probability functions. Belief states are then comparable, and have upper and lower bounds (Bradley, 2016). This is only one way that imprecise probabilities may be formally represented. Other versions assign intervals of probabilities that only specify bounds on degrees of belief, rather than a set of degrees of belief (see, e.g., Kyburg Jr and Teng (2001)). Below, I assume a characterisation in terms of sets of probabilities, and will talk in terms of belief sets. Tying these versions together is a rejection of *precise probabilism*, the idea that beliefs should always be represented by a probability function that takes each proposition an agent entertains to a real number in the unit interval. The application of the equivocation norm assumes precise probabilism, as it always should result in the selection of a precise degree of belief, even if it is determined pragmatically from a set of sufficiently equivocal functions. But as there are concerns about the extent of choice involved in applying such a norm, one might be tempted to be an imprecise probabilist. The norms one applies to constrain belief will therefore depend on the benefits and drawbacks of the imprecise model of belief. Additionally, the four distinctions cut across one another, and the relationship between them is summed up in Table 7.1.

Preferring an imprecise subjective Bayesianism as the way to formally model EEMM would run into problems of subjective choice. From those degrees of belief consistent

	Subjective	Objective
Imprecise	Agents can have different belief sets when faced with the same evidence	There is always a single correct imprecise belief set that a person should hold in light of a given body of data
Precise	Agents can have different degrees of belief when faced with the same evidence	There is always a single correct degree of belief that a person should hold in light of a given body of data

Table 7.1: Table summarising the kinds of Bayesianism along the distinctions between subjective and objective Bayesianism, and precise and imprecise probabilism.

with the evidence, one is free to choose any belief set. However, one might instead be an *imprecise objectivist*. Joyce (2011a) identifies this possibility, although he also notes that there are no accepted norms for selecting a single correct belief set. One option is to implement the *Chance Grounding Thesis* (CGT) as a norm on permissible degrees of belief (White, 2010). This thesis holds that only on known objective chances should we hold sharp credences, otherwise we should spread our credence over the remaining chance hypotheses consistent with our evidence. In other words, permissible degrees of belief are constrained to the entire belief set consistent with the evidence. As Williamson (2010, p.69) identifies, an imprecise version of calibrated Bayesianism is more cautious than his objective Bayesianism, making it better motivated on that count. This imprecise version selects the interval compatible with evidence, so is the same as the imprecise objectivism outlined here. As a result, it is both better motivated, and at no more risk of subjectivity than precise objectivism. There is however more to why one would prefer a model in general than caution and non-evidential norms. To be able to make a full judgement on the most appropriate model to apply to EEMM, more must be said about the relative merits of imprecise versus precise probabilism.

One important motivation is the supposed ability of the imprecise model to offer a better representation of the weight of evidence than the precise model (see §5.3 for details of the distinction between the weight and balance of evidence). There are two ways of explaining this: one in terms of the brute fact of being able to represent weight of evidence in one's credences; and another in terms of decision making. With regards to the first way, Joyce (2005) claims an imprecise model of belief is the only model that can appropriately capture the distinction. This can be illustrated by considering a coin that has unknown bias but has not yet been tossed, and another coin known to be fair that has been tossed many times. The precise probabilist would be forced to assign a point probability to the proposition that the next toss is heads. For both coins this probability would be (approximately) 0.5: the 'unknown bias' coin because of symmetry considerations; the 'known bias' coin because of frequency data. But the imprecise probabilist would say that the probability should be spread over [0,1] for the unknown bias coin, and be 0.5 for the known bias coin. In this way the imprecise probabilist is able to represent the fact that before the coin is tossed our evidence has no or little weight. Weight of evidence is in this model represented by the concentration of probabilities over possible chance hypotheses, and the tendency of these probabilities to change in the light of new evidence. The precise probabilist struggles to represent this dimension of the strength of evidence: the probability of heads is the same for no coin tosses as it is for 100.

The second way of explaining the importance of representing weight of evidence appeals to ambiguity aversion (Bradley, 2016). This is a phenomenon empirically observed in participants undertaking certain decision problems, which shows that agents prefer risky prospects to ambiguous ones. Outcomes of a decision are risky when the outcome is uncertain but occurs with known probability. Ambiguous outcomes are when the outcome occurs with uncertain probability. In the coin tossing example, we might prefer a bet on the outcome of the known bias coin. And this preference can be rationalised by the ambiguity aversion principle, but only when an imprecise model is used. This is because the imprecise model represents one's belief in the unknown bias coin using the maximally ambiguous P=[0,1], whereas the precise model represents both coins with the same probability. So one cannot tell whether the bet is actually risky or ambiguous on the precise model, as both bets are risky. Yet people still avoid bets that are ambiguous. Only the imprecise model can explain this.

Proponents of precise probabilities contest the point that the ability of the imprecise model to capture the weight of evidence is a benefit. Williamson (2010, p.72) argues that in this case the imprecise model conflates evidence with belief. The idea is that we should not be using our representation of belief to model evidence. Rationalising ambiguity aversion by appealing to the model of belief is wrong headed to the precise probabilist. Instead we should have a separate model of evidence which along with the model of belief informs decisions. People then avoid ambiguous outcomes by looking at the evidence. In the coin example, one would prefer a bet on a known bias coin because it has been flipped many times, and not because one's confidence is represented as 0.5 rather than [0,1]. More generally, our model of evidence will tell us whether the degree of belief we hold should be stable or not in the light of new evidence. However, identifying the weight of evidence is still important. And as we have seen, it is also important to represent the stability of one's confidence, in order to more accurately capture and communicate the separate dimensions along which evidence can be strong or weak. It may be conceptually unsound to represent the weight of evidence *solely* in one's beliefs, but it is still appropriate to represent how the weight of evidence is *reflected* in our beliefs. This is because the nature of our beliefs will inform how we act. A model of evidence is less appropriate for this role than modelling the stability of our confidence, and imprecise probabilities have a means to do so baked into the theory. Precise probabilists can appeal to notions of the *resiliency* of probabilities to change (see, e.g., Skyrms (1977)). But they have no natural way to model this in the representation of one's beliefs. So in cases where we want to acknowledge that our degrees of belief can be unstable, the imprecise probability account may be better suited.

Precise probabilists can argue in turn that they have a better representation of the balance of evidence, namely the point valued probability function. One problem with the imprecise representation of the balance of evidence is that belief sets are more difficult to compare than point probabilities. Two belief sets with different bounds may overlap and there is no consensus on which set has the greater or lesser balance. For example, for belief sets $\mathbb{B}_1 = [0.4, 0.6]$ and $\mathbb{B}_2 = [0.3, 0.7]$, one might be tempted to say that $\mathbb{B}_2$ has the greater balance; one might justify this by appeal to the higher upper bound of $\mathbb{B}_2$. But this would appeal to a point probability. One also has to take account of the fact that some of $\mathbb{B}_2$ is lower than $\mathbb{B}_1$. So it is not clear which belief set has the greater balance. This is related to another problem that has to do with the ability to make decisions. To make a decision on the basis of our evidence, we might specify that our degree of belief must be over some threshold τ to trigger an action (Williamson, 2010, p.70). There will be some cases where the belief state overlaps the decision threshold, and there is a question of what to do in such a case. For example, using $\mathbb{B}_1$ and $\mathbb{B}_2$ from above, for $\tau = 0.35$: $\mathbb{B}_1 > \tau$, but $\mathbb{B}_2$ straddles τ. One can arguably trigger a decision when $\mathbb{B}_1$ obtains, but it is unclear what to do in the case where $\mathbb{B}_2$ obtains. A number of approaches have been proposed, but the precise probabilist has a unified approach where precise probabilities are either less, greater or equal than any threshold. A stronger version of this problem is where we have no evidence to constrain probabilities (Williamson, 2010, p.70). In this case, the imprecise probabilist spreads their degrees of belief over [0,1]. But we still might want to make a decision. Such a belief state cannot be greater than or equal to a threshold and so the imprecise probabilist can only suspend judgement, whereas the precise probabilist can make a decision.

This brings to a close my discussion of the details, benefits and drawbacks of competing formal models of belief. In the next section I apply these models to evidence evaluation on EEMM. I will then analyse the appropriateness of each model for this application relative to the model's general benefits and drawbacks.

7.3 Reducing malleability through formalisation

In this section I do two things. One, is to apply the theories of Bayesian epistemology to EEMM. Relative to the benefits and drawbacks of each kind of model, the needs of the application are used to pragmatically determine the most appropriate model of belief for formalising the process. Ultimately, this will be a precise objective Bayesian model. The second, is to consider whether formal models of belief can aid in reducing the malleability of EEMM. I argue that an evaluation that uses calibration and non-evidential norms will restrict subjectivity to a greater extent than one that does not. A problem with this claim is that selection and application of these norms still involves subjective choice. I respond by acknowledging that it is the *extent of relative subjectivity* on a case-by-case basis that matters. Consequently, I recommend that the extent of relative subjectivity required to carry out an evidence evaluation should itself be evaluated alongside the evidence evaluation. First, I introduce an extant formal interpretation of EEMM.

The authors of EEMM supply a probabilistic interpretation, which can be found in appendix B to EEMM (online only, (Parkkinen et al., 2018a)). The bulk of the process does not make reference to this interpretation so it is clear that the qualitative version is the preferred route. While developing my arguments for what formal model to apply to the process, it will help to compare such models with the extant probabilistic interpretation. In short: $x = P(C|E)$ is the degree of confidence in a claim, where E is current evidence, and C is a causal claim; $x' = \sum_{E'} P(E')P(C|E')$ is one's degree of confidence E', where E' are hypothetically expanded future evidence bases; and, $P(x' \in R(x)) \in [a, b]$ is a level of quality, where $R(y)$ is a small interval around y for any y in the unit interval, and $[a, b]$ is some sub-interval of the unit interval. How high the level of quality is depends on the bounds of $[a, b]$. Statuses are again combinations of degrees of confidence and quality levels, such that probabilistic statuses are of the form $x \in [c, d]$ and $P(x' \in R(x)) \in [a, b]$, where $[c, d]$ is some sub-interval in the unit interval and does not have to, but can differ from $[a, b]$. For example, an arguable status may be: $R(1) > x > 1/2$ and evidence is at least moderate quality, $P\left(x' \in\ R(x)\right) > 1/2$.

Note that this interpretation makes use of the concept of hypothetically expanded future evidence bases. Degrees of confidence in causal claims are to be worked out conditional on all possible future evidence bases. What evidence bases are possible is worked out by evaluating the quality of the evidence. Although this is not fully developed in EEMM, this is plausibly carried out by identifying inadequacies in the evidence base, and then working out one's confidence in a claim conditional on an evidence base that did not contain those inadequacies.

7.3.1 Application

Norms for constraining belief

The first thing to consider is what norms for constraining belief would be suitable to apply to the process. At a minimum the appropriate model will apply a calibration norm. The evaluation process uses evidence to constrain belief, and to reduce subjective choice, one should require those beliefs to be consistent with evidence. Before setting out whether this would reduce subjectivity, it is worth illustrating how calibration would work.

To carry out calibration we must decide what we are calibrating to. In the previous section, I put forward the idea that we are to calibrate to physical chance. What would this mean in a concrete application? One option is to calibrate to the p-value calculated in a statistical significance test. The p-value is the probability that the effect size observed in the experiment is due to random error rather than there being a true effect. But this would not tell us the probability of a claim being true. Instead, we might say that we have degree of belief equal to $(1 - p)$ that the observed effect is true. For example, when analysing the amount of virus in lung tissue of rhesus macaques infected with MERS-CoV and then treated with combination therapy, Falzarano et al. (2013a, p.1314) found that:

> "the mean viral load in treated animals was 0.81 log lower than in untreated animals ($p = 0.0428$, unpaired t-test, one tail), demonstrating a statistically significant reduction in virus replication throughout the lung."

In this case one might be tempted to calibrate one's degree of belief b in a hypothesis H about inhibition of viral replication to $(1-p)$: $b(H)$=0.9682. Those familiar with p-values will be aware that there are many common misunderstandings of those concepts. One such misunderstanding is to see p-values as designating the probability of your hypothesis being false, rather than the probability of rejecting the null hypothesis on the assumption that it is true. So one might think taking $(1-p)$ to be the probability that the hypothesis is true is equally wrong. However, if we are calibrating our degrees of belief to the p-value, this is not a problem. Bayesian probabilities need not make reference to repeated applications nor to the rejection of null hypotheses. In this case, we are using frequentist probabilities to assign degrees of belief, rather than incorrectly interpreting concepts that utilise those kinds of probabilities.

One difficulty with this fairly simple approach is that data from mechanistic studies is not always analysed using statistical significance tests. For example, in both studies that investigated *in vitro* inhibition of viral replication (Morgenstern et al., 2005; Falzarano et al., 2013b), inferences are made on the basis of summary statistics alone. In Falzarano et al. (2013b), the conclusions are supported by the observation of a decrease in TCID_{50}, which is a statistic used to measure viral load by calculating the concentration of drug needed to reduce infected cells by 50% (Darling et al., 1998). In this case, one might opt to calibrate to the proportion of infected cells at a clinically relevant concentration. This would however only ever result in a degree of belief of 0.5. Another approach might look at summary statistics and derive probabilities from them. Summary statistics present the information contained in the data as a more easily manageable number, e.g., the mean of the individual measurements in a data set. In Falzarano et al. (2013a), viral load is still measured by TCID_{50}, but what matters is the (mean) *log-reduction* of viral load. Absolute numbers of virus in a sample will be very large, so measurements of reduction in viral load are expressed on a logarithmic scale. Crucially, from a log-reduction number, one can work out a percentage reduction in viral load using the equation: $P = (1 - 10^{-L})$ x 100, where P is percent reduction and L is log reduction. So when $L = 0.81$, $P = 0.84$. As P is a

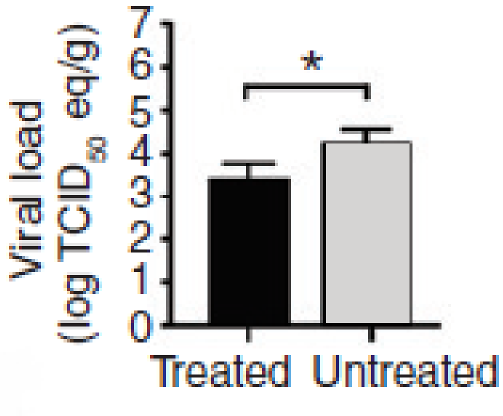


Figure 7.1: Graphical comparison of mean viral load from lung tissues of treated and untreated macaques from Falzarano et al. (2013a). Note the error-bars extending from the top of each bar. I assume that the error is bi-directional, as it is not clear whether the authors considered the error to be uni-directional in practice.

proportion, one might take it to be evidence of physical chance and calibrate to it, such that $b = 0.84$.

In this example, the evidence that one calibrates to is put in terms of a point probability. This does not mean the evidence has determined a degree of belief. Instead, the 0.81 log-reduction is a summary statistic, as it is the mean viral load. But there is likely to be variability in the measured values of viral load that the mean was calculated from. Indeed, in the reports, there are error bars on the graphic representation of viral load reduction (see fig. 7.1). This indicates variability in the log-reduction of viral load that we could convert to variability in probabilities. For example, a visual assessment of the error bars seems to justify a variability of 0.5 log either side of the sample mean (0.81), giving a log-reduction interval of [0.31,1.31]. Using the equation above, this would give an interval of probabilities of [0.51,0.95]. This provides a first approximation to a representation of the stability of confidence: the set of values that are compatible from within the evidence. Of course, this is a simplification of a more complex procedure. For example, the method for calculating confidence intervals may result in different values. However, it suffices for illustration here, and a more complete probabilistic analysis would only improve the accuracy and reliability of the values the interval takes.

The proposal above is only a first approximation, as the evaluative process must capture the ability of an evidence assessment to reduce or increase stability of confidence. If one finds an inadequacy in the evidence base, it loses informativeness and one's stability of confidence decreases. So a full modelling of stability should be able to accommodate this process as well. In the case of Falzarano et al. (2013a), the main problem with the study was one of system dissimilarity, namely, that rhesus macaques only exhibit a transient nature of disease. de Wit et al. (2013) characterised the rhesus macaque model of MERS and found an approximate 1-log reduction in tracheal viral load between untreated animals at 3 and 6 days post infection. Falzarano et al. (2013a), on the other hand, found a 2-log reduction in tracheal viral load in treated animals. We might infer from this that a 1-log reduction of viral load is attributable to the transient nature of disease. It is plausible then that a marmoset model would result in reduction in viral load 1-log less than observed in a macaque model. Changes by 1-log for all values over 1-log, i.e., the difference between 1-log and 2-log, 2-log and 3-log, etc., correspond to a 0.09 percent change. So we might say that a marmoset model would result in a change in degree of belief 0.09 in the negative direction. One cannot say that our actual degrees of belief should change this much, as the evidence for it is indirect. But the probability interval worked out above might extend from [0.51,0.95] to [0.42,0.95]. This is a simple example, and again a more sophisticated analysis is likely to engender more precise assignments of stability of confidence.

From within this set, one might select 0.84 to calibrate one's degree of belief to, due to it being the probability corresponding to the sample mean of 0.81-log. The objective Bayesian would instead equivocate and select a degree of belief from those compatible with the evidence that is sufficiently close to the equivocator function. Motivating the choice of equivocating, rather than taking the mid-point of the interval to calibrate to, are again pragmatic considerations. On average one is less likely to make risky decisions on the equivocation approach than the mid-point approach. A τ greater than 0.5 is often the result of an imbalance of benefits and harms in favour of harms. So there will be many cases where the mid-point triggers decisions, where the equivocation approach does not. In the running example, if τ were to be set at 0.8, this would imply that there must be clear evidence of benefit to trigger a decision. The mid-point approach would trigger a decision, whereas the equivocation approach would not. The equivocation approach is arguably more motivated here as it would not mandate making what is a fairly risky decision. For this reason I utilise the equivocation approach from here on.

In the running example, it is not however clear whether one's degree of confidence should be approximate to 0.51 or 0.5, corresponding to selection from either [0.51,0.95] or [0.42,0.95]. First note that this value is approximate as I leave open the option that the sufficiently equivocal degree of belief is worked out pragmatically from the set of sufficiently equivocal belief functions. Next note that the original formulation of objective Bayesianism does not take into account the way a reduction in the weight of evidence may extend the set of evidentially compatible probability functions. One might think that as the degree of belief is intended to be one's actual degree of belief, then it should only be selected from the *original interval* [0.51,0.95], which is derived from the data, rather than the *extended interval* [0.42,0.95], which is an extent of probability functions that are merely *possible*. To be more in accord with the original formulation of objective Bayesianism, I will assume from here on that the original interval option is correct, such that in the example above, one would select $b \approx 0.51$. The extended interval option is however still a distinct possibility, and it would be important to develop a correct approach. In this case, the difference in degrees of belief selected from each interval is only 0.01. But if, for example, the original interval is [0.7,0.9] and an extended interval is [0.5,0.9], then the competing proposals would select $b \approx 0.7$ or $b \approx 0.5$, respectively. So there is the possibility of significant divergence in the two approaches. My arguments below would hold on either approach, but it is clear that this is an avenue worth exploring in future.

Applying some non-evidential norm to select a degree of belief reduces subjectivity. Without some norm for selecting degrees of belief, evaluators could use their expertise to *elicit* their degrees of belief from within those that are compatible with the evidence. But this would leave status judgements open to accusations of being influenced by subjective choice. So applying some non-evidential norm seems better motivated than just calibrating, when the purpose of applying a formal model to EEMM is to ameliorate malleability concerns. Applying both calibration and a non-evidential norm should reduce subjective choice more than an evaluation that applies neither. An evaluator using the qualitative version may end up not using evidence of physical chance to set their degrees of confidence in a claim. It is consistent with the guidance that they could choose any degree of confidence that is consistent with the evidence. For evidence that is in favour of a hypothesis this could plausibly be any value above 0.5. In my running example, this seems to introduce little inconsistency, given that degrees of belief above 0.5 are still possible. But it is not so hard to imagine a case where applying a calibration norm restricts degrees of belief to an interval I where $0.75 \geq I > 1$, so is qualitatively in favour of the claim. Without a stipulation to calibrate, an evaluator may choose, relative to their expertise that they are at least 50% confident in the claim. When a non-evidential norm is also applied, we see that subjective choice would be constrained further still. Two evaluators could easily differ on what degree of confidence this evidence should entail: one might hold moderate, or one might hold high confidence. This might be because the evaluators differ in their background knowledge. And this would affect the status judgement as it could prove the difference between a ‘provisionally established’ and ‘arguable’ status. A calibrated model would restrict their choice to evidence of physical chance, leading to less likelihood of disagreement. A non-evidential norm would then restrict their choice even further.

One objection might be made to the fact that in my example above, one can calibrate to both p-values and summary statistics. So in this one study there is a conflict between which element of one’s evidence of physical chance to calibrate to. A reason for opting for the summary statistic might be that it is a more accurate estimation of the extent with which virus replication is inhibited. The p-value is the probability that the results are due to random error rather than there being a true effect, and we have interpreted (1-p) as the probability that the effect observed is true. But, we are evaluating a mechanism claim about the effectiveness of combination therapy, so it might be more relevant to calibrate to the extent to which the intervention is effective at reducing viral load. We might, for example, have observed a statistically significant reduction of viral load that equated to only a 20% reduction, and it would seem odd to say we are 95% confident that the mechanism obtains on the basis of this evidence. The problem with this line of reasoning is that there are many points at which subjective choices influence the outcome. Because of this, it might still be the case that calibration results in the admission of too much subjective choice. That a model that calibrates is less subjective than a non-formal evaluation may not convince someone who is worried about the malleability of EEMM in general. I defer an answer to this problem until the end of the section, as a fuller response will build upon the arguments I make below. In the remainder of this section I will however use the approach that utilises summary information (such that $b = 0.84$) as an extended example, varying it where appropriate to illustrate how one would apply other kinds of formal model. The next question to ask then is whether this kind of model is most appropriate for EEMM, relative to the needs of the process.

I have said that the most appropriate model for EEMM would be determined relative to the needs of the process. Thus far, I have provided examples of how a precise, objective Bayesian model would work to constrain belief and how it reduces subjective choice. In favour of this model is that it broadly fits with the way EEMM models confidence: split into a degree of confidence and an extent of the stability of that confidence. The examples above also do this, e.g., if one uses an equivocation norm, one might hold degree of belief in the claim approximate to 0.51, and the extent of stability of confidence is modelled by [0.42,0.95]. However, the probabilistic interpretation of EEMM models one's stability of confidence as holding point valued beliefs constrained by hypothetical evidence bases, separate to those constrained by the current evidence base. In either case, one would be selecting point valued degrees of belief. But my proposal seems more straightforward to implement. The EEMM approach requires one to: i) select a possible degree of confidence in a claim on a future evidence base; ii) work out another degree of confidence for how close the degrees of confidence in the claim are on current and future evidence bases.

Calculating this extra probability value makes the EEMM interpretation more complex. My approach only requires calculating a range of degrees of confidence possible on a future evidence base and a degree of confidence. Moreover, it is not clear how one would go about calculating a probability for the closeness of the current and future degrees of confidence. By contrast, there are tools available for selecting intervals of probabilities from data. My approach is thus more direct as it derives stability from probabilistic information provided by the evidence. For that reason, I assume below that my approach is the correct way to model stability. It should however be noted that neither approach is near full development as yet.

Precision of degrees of belief

One might object to the claim that a model that applies both calibration and equivocation norms is best suited for applying to EEMM on the grounds that the process appears to be constructed with a precise model in mind. The problem is that EEMM assumes that one should always select precise degrees of confidence, but offers no argument for this position. It is also the case that I have offered no argument for applying only a precise model to EEMM. I have assumed that an equivocation norm is the right way to reduce subjectivity, but a proponent of the imprecise model might insist that to adjudicate on what the most appropriate model is, requires considering the possibility of an imprecise EEMM alternative. To more completely assess appropriateness, I intend here to sketch one, and contrast this with the precise model I have sketched out above.

An issue the imprecise model might face is representating the balance of evidence. Indeed, Joyce (2005) attempts to fashion an imprecise representation of balance, but admits there is no full development. The precise model has an intuitive way to represent balance, which is exemplified in the probabilistic interpretation of EEMM, where x stands for degree of confidence. A proposal for overcoming this difficulty would be to select a set of probability functions compatible with one's evidence of physical chance. This idea makes use of the chance grounding thesis, which was introduced in §7.2.2. In the running example, one might calibrate one's belief state $\mathbb{B}$ to the interval of probabilities calculated from the log-reduction information: $\mathbb{B} = [0.51, 0.95]$. But, the imprecise model represents stability as the extent of the 'concentration of probabilities' Joyce (2005). So the initial belief set, which typically represents the weight of evidence on the imprecise model, would also represent the balance of the evidence in this imprecise EEMM alternative. To overcome this conflict, we would need some way of also representing the weight of evidence. A possible way to do this is to introduce a second interval I', where $I \subset I'$, and I' is in the unit interval. The extent of stability can be represented as the size of the set $\mathbb{P}$ of probability functions p such that $\mathbb{P} = \{p : p \in I' \wedge p \notin I\}$. This means that $\mathbb{P}$ consists of all p that are possible given a future evidence base. In the example, for original interval $I = [0.51, 0.95]$ and adjusted interval $I' = [0.42, 0.95]$, $I \subset I'$ and the extent of stability of confidence is $|\mathbb{P}| = [0.42, 0.51)$.

The problem with this proposal comes down to decision making. Recall that the general decision making criticism of the imprecise model was that whether we decide to act depends on our beliefs crossing some threshold τ, and when our beliefs are represented by some interval that straddles τ, then we face difficulties in triggering decisions. On EEMM, statuses are important for making decisions. If the degree of confidence element of a status is an imprecise belief state $\mathbb{B}$, then there may be cases where $\mathbb{B}$ straddles some trigger threshold. Plausibly, this problem transfers to decision making on an imprecise EEMM. In the extended example, evidence for viral load reduction constrains degrees of belief to an interval $I = [0.51, 0.95]$. On the imprecise model, $\mathbb{B}$ = [0.51,0.95], and stability is modelled as [0.42,0.95], with extent $|\mathbb{P}|$. On the precise (objective Bayesian) model, one's stability of belief is modelled (at least initially) as [0.42,0.95] with degree of belief $b \approx 0.51$. Suppose that the trigger τ for recommending combination therapy is 0.75. As it is the status of the causal claim that matters for decisions, the degree of confidence in this status must be greater than or equal to 0.75 to recommend the treatment. Suppose further then that one's degree of confidence in the mechanism claim is calibrated to b, and

b is less than one's degree of confidence in the correlation claim (this final supposition is not strictly true, but this example is for illustrative purposes). So the degree of confidence in the causal claim also equals b. The precise model can make a decision: if $b \approx 0.51$ then a decision is not triggered. Arguably, the imprecise model cannot: $\mathbb{B}$ straddles τ so it is unclear whether a decision should be made or not. So the imprecise model is not suitable for use in the evaluation process as it cannot help to make decisions.

A proponent of the imprecise model might object that when intervals straddle thresholds, one should suspend judgement. It is no problem, so this argument goes, that one cannot make a decision in such cases, as decisions are not warranted due to the unspecific nature of our evidence. The precise model gets it wrong by making decisions in cases where we should not. One response to this might note that there is a pragmatic need to make decisions. For example, there are no recommended treatments for MERS and anything that might work is worth making a decision on. But the imprecise model cannot make a decision, so the precise model is more appropriate. One might think to the contrary that we should not be basing decisions on unspecific evidence, even if there is a demand for treatments. Further, making no judgement until we acquire more evidence should be motivated by the low specificity of the evidence.

However, on the imprecise model, indeterminate belief states are possible that are not the result of highly unspecific evidence. And in those cases, one's belief state may still straddle a threshold for decision making. For illustration, consider a plausible case where one's evidence is quite weighty, such that the initial balance on an imprecise model restricts $\mathbb{B}$ to $I = [0.8,0.9]$, stability is modelled by $I' =[0.78,0.92]$, and $\tau = 0.85$. So in such a case $\mathbb{B}$ straddles τ and the proponent of the imprecise model will struggle to say what sort of decision should be made here. Again, they might call for suspension of judgement. However, suspension of judgement is arguably not a warranted attitude to take towards this evidence. I is fairly precise, and I' indicates that this evidence is beset by few inadequacies. The weight of the evidence must then be fairly high. As a consequence, the call to suspend judgement depends entirely on the value of τ, rather than both τ and the strength of the evidence. On the other hand, a precise model would be able to make a decision, and would do so through a process that is transparent and contestable. For example, an objective Bayesian model would not recommend the treatment as one would select a degree of belief closest to the equivocator, which will be less than 0.85. And this decision would be made on the basis of consideration of both τ and the strength of the evidence. So the imprecise model again suffers in a decision making context, and one is not warranted to suspend judgement.

Another worry about appealing to decision making to support the precise approach, is that it may suffer from the same problem the imprecise approach does. This is because the precise approach still needs to recognise the weight of evidence, otherwise decisions may be triggered on bad evidence. To do so would require considering stability of confidence when making decisions. But, in the example above, the decision is triggered solely on the basis of the point valued degree of belief. The problem is that the range of possible values [0.51,0.95] includes many values that are also above the threshold. One might require instead that to make a decision requires the extent of stability to be entirely above the trigger level. The problem with this is that when one takes into account the stability of confidence, the precise model and the imprecise model might seem to have the same difficulty in making decisions: the extent of stability of confidence on the precise model straddles τ. Moreover, in those cases where the precise model could make decisions, the imprecise model could as well: only when the lower bound of $\mathbb{B}$ is greater than or equal to τ can one trigger a decision. So on the count of decision making, there would seem not to be much difference between the two models.

One might think that one need not include the range of possible values in a decision making process because the motivation for non-evidential norms carries over to the decision making process. In the example above, Equivocation selects a degree of belief below τ and the application of the norm is motivated by caution. Arguably, not triggering an action in such a case is also motivated by pragmatic concerns about caution. When τ is a trigger level for recommending an intervention, a high level for τ indicates that we need strong evidence of benefit. This could be because benefits do not strongly outweigh harms for that intervention. So it seems prudent to not trigger actions in cases where an interval straddles τ. This intuition is captured by a precise approach that applies Equivocation. Importantly, the stability of confidence, and so the weight of evidence, is taken into account during the decision making process as it represents a range of values that are possible. A degree of confidence is then selected from within this range. Plausibly, if we have a wide interval then we are less likely to trigger a decision.

However, this might not be the case when τ is low. Consider a case where $I = [0.2, 0.6]$ and $\tau = 0.4$. Application of Equivocation would select $b \approx 0.5$, so the action would be triggered. For an intervention, such a case indicates that benefits do fairly strongly outweigh harms. Caution still reigns here, as even if the interval straddles the threshold it might be more cautious to trigger a decision, given that the intervention is more likely to have a benefit than a harm. One might worry that this means stability of confidence is not playing much of a role in the decision making process. In this case, the interval straddles τ, yet we still make a decision, which is inconsistent with the example above. The problem is that for similarly unstable beliefs, it is again the place of the threshold that ensures whether a decision is made or not, rather than the extent of stability of confidence. In the low threshold case, some possible degrees of belief are also below τ. While the risk profile of this intervention may be low, it is not non-existent. A τ of 0.4 indicates there are *some* harms. Given that the confidence is fairly unstable on the evidence for this intervention then it seems wrong to trigger a decision here. Therefore, caution does not motivate triggering an action in the low-threshold case, even if it does motivate not-triggering an action in the high-threshold case.

The resolution of this problem starts with noting that thus far I have considered integration of stability of confidence into the decision making process as a matter of where the range of possible values fall, i.e., whether having the lower and upper bound of the interval on different sides of the threshold is a problem. This is not the only option. Just as in the example from my response to the suspension of judgement problem, there will be some cases where we have good evidence, but the interval still straddles τ. In the low threshold case from the previous paragraph, the interval is quite wide, making many values below the threshold possible. Consider an alternative case where $I = [0.35, 0.45]$ and $\tau = 0.4$. Equivocation would select $b \approx 0.45$, thus triggering the decision. Triggering a decision in this case seems much less imprudent than in the example from the previous paragraph. And this is due to the interval being quite narrow, rather than any change in the threshold. In other words, the point probability is supported by good evidence even if the interval straddles the threshold. Comparison of the two low-threshold examples motivates the implementation of an 'extent of stability' approach to decision making. The intuition that one should make a decision in the narrower interval case, and not in the wider interval case, is motivated by the recognition that it is the extent of stability of one's beliefs that factors into decision making, rather than the point values that are possible. So a decision making process should be grounded on: i) whether the point probability is over/under threshold; ii) the extent of the stability of that confidence. Of course, this proposal needs further development. One issue is that it may run into arbitrariness problems when one has to mandate a point at which one has stable enough beliefs. In keeping with the arguments of this chapter, one could reasonably determine this pragmatically. But if we are to preserve the idea that decisions can be made on narrow intervals, even when they cross a decision-making threshold, then an extent of stability approach is the right way to go if we are to also include stability of confidence in decision making.

In sum, any model that calibrates to empirical evidence will restrict subjective choice greater than a qualitative model of assessment on EEMM. Applying a further, non-evidential norm, along the lines of an objective Bayesian epistemology would seem to restrict subjectivity even further. Applying such a norm would assume a precise probability model. An imprecise model would restrict subjectivity to the extent that it calibrates to empirical evidence and supplies some motivation for the selection of a particular belief set from those compatible with the evidence. However, the imprecise model struggles with decision making. One consequence of this is that the imprecise model is less apt for application to EEMM, given that decisions are to be made on the statuses of claims. Therefore, applying a precise model to EEMM is better motivated. Further, selecting a precise model should lead to one applying norms along the lines of objective Bayesian epistemology.

7.3.2 Suitability

There are three inter-related problems with utilising an objective Bayesian model to ameliorate malleability. One is that the arguments above assume that precise degrees of belief can always be selected, when in reality, the practical application of the norms may be difficult. There are two separate issues with application here. One is that selecting precise degrees of belief may involve an unacceptable degree of *idealisation*. Imprecise probabilists criticise the precise approach because it is unrealistic: people rarely if ever hold such specific beliefs about the propositions they entertain. Part of the reason for this is that one's evidence is often unspecific, and so the proper response to such evidence is to hold similarly unspecific beliefs. The precise probabilist would say that one *should* select precise degrees of belief, even if in practice one cannot always do so. In other words, such models are normative in nature, rather than descriptive. However, because they are not always realistic, the selection of precise degrees of belief in practice may depend on arbitrary choices. So any attempt to ensure better calibration to the 'true' physical chance is fruitless, as subjective choice can always bias the final judgement. The other potential issue with selecting precise degrees of belief is with the application of the calibration and equivocation norms. Practically, they may be difficult to implement, which means again resorting to idealisations. The other problem with application is that both calibration and equivocation require many choices to be made in their application. For example, I identified above that one could calibrate to p-values, or to summary information, which resulted in either b=0.9468 or b=0.84, respectively. So choices at this point will affect what degree of confidence one holds. Overall, the application of formal models still requires subjective choices to be made. Training evaluators on these methods, or including experts in data analysis on evaluation panels, may ameliorate some practical difficulties. But even if that were the case, one is still faced with the issue of subjective choice.

My resolution to the problem of applying norms in practice is to suggest a compromise. The basic idea would be for evaluations to utilise formal tools that approximate as far as is practically possible an ideal model, in this case, precise objective Bayesianism. This idea can be illustrated by consideration of different contexts that determine the extent of the applicability of formal models. For example, it might be the case that one can use the tools of the imprecise model to guide reasoning without having to select a precise probability. This might be because the decision making context is one where imprecise belief states do not hinder making a decision. A simple case is where for trigger level τ, the entire interval selected by the evidence is greater than τ. Equally, one may have to act *as if* a precise probability can be selected if the decision making context requires it. If the practical context makes using the tools of objective Bayesianism difficult, then evaluators would have to rely on their judgement to pick out a degree of belief from within those consistent with the evidence. In general, the choice of which elements of a tool to apply would depend on context, but the ideal to approximate would be a precise objective Bayesian theory. The norms of this model would guide evaluator reasoning during an evaluation. Further, it would allow confidence to be constrained as much as is practically possible within the operating context. In general, instead of seeing formal models as a requirement that must always be implemented, they should be seen as a set of tools that may or may not be applicable per context.

Of course, how to approximate models will require making many choices. And approximation does not deal with the problem of the choices that must be made during the application of the norms. This relates to the third problem: how one decides which is the right non-evidential norm to select. I preferred Equivocation, but there may be other norms for selecting degrees of belief. For example, above I considered the option for calibrating to the mean. One could attempt to motivate this through non-evidential reasons. In the case of Equivocation, one would appeal to caution. But, again, one might worry that this means the selection of degrees of confidence is influenced by subjective choice.

A response to this problem depends on recognising rather than ignoring where subjective choices influence decisions. What matters is the extent of *relative* subjective influence. There are choices that must be made at every point in any inference process throughout science. For example, in the investigation into the existence of the Higgs Boson in High Energy Physics (HEP), there has been controversy over the particular α-level that p-values must be less than for results to be deemed statistically significant (Staley, 2017). As noted above, p-values are interpreted as physical probabilities. Moreover, HEP is an established science and experiments within this field can draw on sample sizes of data that dwarf those found in medicine. One might be tempted to claim that the inferences made in this example are (near to) completely agent-independent. However, choice of the appropriate α-level has to be argued for, and Staley argues that it comes down to a consideration of risk. In that case, the extent of the influence of subjective choice on the conclusions is minimal, but is still present. On the other hand, a criticism of Stegenga's master argument for medical nihilism was made against the almost purely subjective nature of the probabilities used in the calculation (§6.3). In a similar vein, the problem with a model such as *E-synthesis* is that the probabilities utilised in it are the subjective probabilities of experts (§7.1). The influence of subjectivity on inferences in those cases is (almost) maximal. But the influence of subjectivity in a model that employs norms that actively seek to constrain subjectivity is likely somewhere in-between. Taking motivation from Sprenger (2018)'s defence of the objectivity of subjective Bayesian inference (see §6.3), if we make the reasoning that goes into the application of certain norms explicit, then the subjective choices that are made are put on display. It seems that there will always be some degree of subjective choice required, and hiding it away is not going to head off malleability concerns.

Being explicit about the extent of subjective choice in each assessment will help. In general, it would be valuable to conduct a separate assessment of the extent of subjective choice in an evaluation, alongside the evaluation of the causal claim itself. This would be an addition to EEMM, or any evaluation framework. In particular, when approximating the ideal model, choices made when one cannot implement the ideal model must be made explicit. One can then decide whether they are reasonable or not. Additionally, the motivations for applying an equivocation norm, for example, are clear and have been argued for at length. Contesting the application of this norm is easier than if one were to select degrees of belief based on expert judgement, reasons for which may be less accessible. So worries about letting subjectivity in at the point of choosing non-evidential norms are alleviated by noting and explicitly evaluating the extent of relative subjectivity on a case-by-case basis.

7.4 Conclusion

In this chapter, I considered formalisation as a way to constrain subjective choice in EEMM. Theories of Bayesian epistemology were contrasted against one another, identifying benefits and drawbacks of adopting each as a model for: 1) imposing constraints on belief; and 2) probabilistically representing one's belief state. I then identified which model is most appropriate for application to EEMM. Model appropriateness was determined pragmatically, matching up the needs of the application with the benefits and drawbacks of each formal model. I concluded that the most appropriate model was an objective Bayesian epistemology that utilised precise probabilities. This is because an objectivist model rules out more subjective choice than other models, and a precise model runs into fewer problems concerning decision making than does the imprecise model. I considered an objection based on whether the use of formal models in general was suitable for restricting malleability in EEMM. I argued that a fully formal model was inappropriate, but should be used as an ideal to approximate and guide reasoning. I suggest that a separate evaluation of the extent of subjective choice required to implement both the formal model and the evaluation process should be carried out alongside the evidence evaluation.

Part IV

Extensions

Chapter 8

Additional guidance

8.1 Introduction

In previous chapters I have dealt with malleability as an abstract concern about meta-methods used for evidence evaluation. To make my claims about routes for restricting the risk of malleability more precise, something more concrete is in order. In this chapter, I use concrete examples from the EEMM process and the review I undertook in Chapter 3, to show how the formal models and reasoning principles introduced in previous chapters can be used to extend the process and restrict malleability.

In §8.2, I identify a part of the EEMM process at risk of influence by subjective choices. This is the point at which evaluators must decide on *relation weightings*, where a relation weighting is made when one decides which evaluative criteria affect which evaluative judgements. For example, a decision must be made on whether a 'status criterion' affects either the degree or stability of one's confidence, or both. At present there is no restriction on what kind of relation weightings can hold, but, given the concepts and principles introduced in Chapters 5 and 7, it is plausible that there should be. I then use those same concepts and principles to formulate some additional guidance on relation weightings in §8.3. To be specific, concepts such as the weight and balance of evidence, and the principles that govern how evidence imposes constraints on belief in Bayesian epistemology, in turn impose restrictions on the kinds of relation weightings possible. I then show in §8.4 how these restrictions work in a re-evaluation of my evaluation of the quality of mechanistic studies and status of a mechanism claim from Chapter 3. This provides a concrete example of how the guidance developed in this chapter improves on the guidance in EEMM. Moreover, it provides an indication of how philosophical and conceptual work from this thesis can be used to extend EEMM.

8.2 Subjective choice and making relation weightings

In the evaluative process of EEMM, *evaluative judgements* are reasoned to by considering *evaluative criteria* (these terms were introduced in §3.2). Decisions must be made about how criteria affect judgements. One kind of decision is made about the differential *extent* of the impact different criteria may have on judgements. For example, an evaluation may find that a study has set no control on a potential error, and used an experimental system that differs from the target system in important ways. But the evaluators may judge that the lack of an important control is a more serious problem than the system dissimilarity. So they will give more *weight* to this detail when judging the impact on their confidence. As they are quality criteria, lack of controls would decrease stability to a greater extent than system dissimilarity does. This is an example of what I call an *extent weighting*. By contrast, when evaluators make *relation weightings* they make a decision about which aspect of confidence is effected by particular criteria. To illustrate what a relation weighting is, consider the status criterion ‘there are many sources of evidence for crucial features’. One might think that presence of evidence for crucial features should raise one's confidence in the claim. This would be to make a kind of relation weighting, where a criterion is deemed to affect degree of confidence. One might also think that doing well on this criterion will stabilise one's confidence in a claim, which is another kind of relation weighting. I return to this example below, once I have formulated guidance on what kinds of relation weightings are possible.

The first thing to note about this distinction between kinds of weightings is that *weighting* and *weight of evidence* may appear to mean the same thing. There are however important differences between them. The weight of evidence concerns the informativeness of an evidence base, and is one dimension of the strength of an evidence base. This means low weight of evidence will lower the strength of evidence, and vice versa. A weighting also concerns informativeness of the evidence. Making a weighting involves making a judgement about the impact features of evidence have on a judgement of the strength of that evidence. But it is not the case that giving a greater weighting to the impact a *criterion* has on confidence necessarily increases the weight of the evidence. Recall from §5.4.1 that identifying inadequacies in an evidence base reduces its informativeness, and consequently its weight. For a criterion that identifies an inadequacy in the evidence, weighting the extent of its impact on confidence as 'high' will therefore highly *decrease* the weight of evidence. So while the direction of the change in strength of evidence depends on the direction of change in the weight of evidence, this is not the case for a high extent weighting. The direction of changes in the strength of evidence will depend instead on whether the criterion identifies an inadequacy or a strength. So the terms cannot be referring to the same thing.

The next, and most important thing to note about relation weightings is that they are a point at which I claim subjective choice has potential to negatively influence judgements. The process is thus at risk of a concrete charge of malleability. The reason why I make this claim is that there are no explicit restrictions on the kinds of relation weightings possible, whereas it is plausible that there should be. Given my arguments linking 'quality' and ‘stability’ in Chapter 5, it seems easy to make a relation weighting between quality criteria and stability of confidence. However, there is no well defined relation between status criteria and either stability or degree of confidence. Recall that a status judgement is composed of both one’s degree of confidence in a claim and the stability of that confidence. The guidance states that status criteria are meant to raise or lower status, but it does not stipulate whether that is through raising or lowering the degree or stability of confidence, or both. The relation weightings possible are left to expert judgement. However, it is plausible that there should be restrictions on the kinds of relation weightings possible. To see why, consider my claims about what makes evidence weighty.

In Chapter 5, I argued that there were features of the evidence that only contribute to the weight of evidence, and so to stability of confidence. For example, the gross amount of information available seems to be reflected in only the stability of one’s confidence, rather than in its extent. This relation is seen in the formal models of belief as well. For example, degrees of belief are calibrated to evidence of physical chance, which is another way of saying this feature of the evidence contributes to the degree of confidence only. So some features of evidence seem to only affect one part of the status judgement. It is plausible that any restrictions on relation weightings should be underpinned by the reasoning concepts and principles that I argued underpin the evaluation process, namely the weight and balance of evidence (§5.3), and norms for imposing constraints on belief (§7.2.2). But there are no explicit restrictions in place to ensure this. The problem is that if relation weightings run counter to these plausible reasoning concepts and principles, then an evaluation ends up losing the benefits of reducing subjectivity accrued through following them. Malleability again becomes a concern.

Making this issue more than just an abstract possibility is the fact that evaluators will be experts in their own particular fields. They should not be expected to have implicit knowledge about the reasoning principles that underpin the evaluation process. In a review on treatments for MERS, the evaluators are likely to be virologists, clinical scientists, or something similar, with a wealth of knowledge in their domains. This will include both theoretical and experimental knowledge, but they cannot be expected to have knowledge of the kind of meta-methodological principles that I argued are required to make judgements about relational weightings. This is why it is plausible, although not certain, that evaluators may make relational weightings that run counter to the epistemic principles that underlie evaluations on EEMM. In general, evaluators cannot be expected to hold this kind of meta-methodological knowledge. It is instead the job of methodologists to provide as much guidance as is possible on these matters. What is needed is provision of guidance on what kinds of relation weightings are possible on EEMM. Before providing such guidance there are some clarifications to be made, and a worry to identify.

Firstly, as noted in the previous paragraph this problem is only plausible, rather than a concrete issue facing the process. I intend to show in §8.4 that it is a concrete problem by comparing reviews that are carried out on a process with and without some guidance on relation weightings. Of course, to do so will require formulating the guidance, so I will carry that task out in the next section. Secondly, even as a plausible problem it does not seem to beset the current formulation of GRADE. However, if GRADE takes on my recommendations from Chapter 5, and switches to a characterisation of quality in terms of stability of confidence, then it may have to consider the problem of making relation weightings. There will have to be some features of evidence that determine degrees, and some that determine stability, of confidence. Until a change is made, not much more can be said about what relational weightings could be made. What I say here may transfer over to GRADE, and can be used as a starting point if in fact GRADE is adjusted in line with my recommendation.

Finally, a worry about trying to provide restrictions on the scope of expert judgement, is that EEMM is intended to be generally applicable across many different domains within medicine, broadly construed. Those domains may differ on the relative importance of certain evaluative criteria. This is one motivation for allowing a significant degree of latitude to evaluators when making judgements. For example, IARC employ their ten key characteristics of carcinogens when evaluating the potential carcinogenicity of compounds (Smith et al., 2016). The characteristics are also compatible with the EEMM approach (Parkkinen et al., 2018b, p.103). EEMM could thus be used by IARC, and the characteristics could help to determine whether a mechanism claim holds. However, if EEMM restricted evaluation to only the process found in the guidance, then this domain specific knowledge would be left out of an IARC evaluation of evidence of mechanism. Such an evaluation would arguably be impoverished. So it is right that there must be some room to make judgements about the application of the process to the specifics of a domain. But it is also the case that restrictions are a part of every evaluative process. For example, GRADE take a similar approach to the need to allow subjective judgement, particularly about extent weightings. And it seems like making extent weightings is one part of an evaluation that would not benefit from restrictions, as they will be highly context sensitive. But GRADE do place restrictions on other parts of the process. For example, evaluators are allowed to include evaluative criteria other than those contained in the explicit framework, but there will be restrictions on what kinds of criteria are permissible. Experimenter hair colour is plausibly not a permissible criterion. At root, all evaluative processes must restrict expert judgement as much as possible, without hindering the evaluation. I do not argue that having room for expert judgement is wrong, it is just that, at least where making relational weightings is concerned, there is too much room in EEMM.

8.3 Guidance on relation weightings

Features of evidence that contribute to its weight, and those that contribute to its balance, can be categorised by whether they impose one of two broad constraints on aspects of confidence. In Chapter 5, I developed the idea that the extent of the informativeness of the evidence, understood in a framework of ruling out alternative explanations, was what made evidence weighty. In developing this idea, it was assumed that what gave evidence its balance was frequency data. This assumption was made on the basis of the way balance is reflected in degrees of belief in the toy examples used to illustrate the distinction, namely coin flips and the x-sphere example. As identified in §7.2.2, frequency data is *one way* of obtaining evidence of physical chance, to which degrees of belief should where possible be calibrated. But evidence of physical chance is not limited to frequency data (see discussion on objective Bayesianism in §7.2.2 for expansion of this point). Therefore, the broad feature of evidence that gives it a particular balance is *evidence of physical chance*. These broad categories are a first approximation to guidance on relation weightings. Criteria that affect stability will be restricted to those that are relevant to *informativeness*, and criteria that affect degree of confidence will be restricted to those that are relevant to *physical chance*. I will distinguish further within those two broad restrictions in the remainder of this chapter.

8.3.1 Informativeness

Broadly, an evidence base that is more informative will result in greater stability of confidence than one that is less informative. Recall that informativeness can be distinguished into *gross* and *net* informativeness, where gross informativeness is the total amount of information, and net informativeness is the total amount of information less any inadequacies in the generation and analysis of the evidence. Those inadequacies explain the results of a study in terms other than that the hypotheses under consideration is true. One might think there should be separate instructions for identifying criteria that relate to gross informativeness and those that relate to net informativeness. In §5.4.1, I posited a number of options for what can contribute to the total amount of information, including but not limited to the number of studies and number of measurements. In the context of an evaluation by EEMM, what contributes to gross informativeness could thus depend on what is being evaluated. If it is a study being evaluated then what matters will likely be the number of measurements (the sample size); equally, if the object of evaluation is the evidence base, then what matters will be the number of studies.

However, it may be difficult to tie the extent of the gross informativeness of an evidence base to stability of confidence. The first difficulty is with identifying some threshold over which evidence has high gross informativeness. For example, when flipping a coin as evidence for the proposition 'the next coin flip will land heads', 1000 coin flips may be requisite for high gross informativeness. But 10,000 coin flips is even more informative, and 100,000 more informative still. There will however be some point at which the coin flips will settle near a stable number for the frequency of heads. This has been proven in practice (e.g., Gillies (2000, Ch.5)) and in computer simulations (e.g., Tijms (2012, Ch.2)). Moreover, in more complex set ups such as scientific experiments, formal tools can help to work out what counts as informative *enough*. There is a close link between sample size and the width of a confidence interval. And there are methods for the calculation of what sample size is sufficient for detecting real effects in each experimental context (Pogue and Yusuf, 1997; Guyatt et al., 2011b). So we could instruct evaluators to identify whether a criterion indicates that the evidence has reached some threshold for having high enough informativeness to make confidence stable.

When we look at how stability of confidence is worked out in the net informativeness framework, it looks like instructing evaluators to consider the gross informativeness of an evidence base is misguided. This is because even if we instruct evaluators to relate effects on stability of confidence to particular criteria on the basis of a threshold of gross informativeness, effects are still dependent on considering alternative explanations. To see why this is so, consider first how the net informativeness framework works, and how one can formulate guidance on relation weightings from it.

The idea of net informativeness relies on the notion of alternative explanations. To make this work in the practice of an EEMM evaluation, the basic idea introduced in §5.4 was combined with the formal models of belief in §7.3.1 to model stability using first, the range of degrees of belief compatible with the evidence, and then, an expanded range of belief obtained by considering alternative explanations. Meeting the threshold for gross informativeness works to increase stability by ruling out alternative explanations in terms of inadequate sample size and/or random error. This means what is being worked out is always net informativeness. One might think that in the case where evidence is maximally informative, gross informativeness might work in isolation from consideration of alternative explanations. But it is doubtful that this is ever the case, and even if it were possible, the alternative explanation framework is how one reasons to having maximally stable beliefs. So the amount of total information is not considered in isolation and what matters for making relation weightings is alternative explanations and net informativeness.

Guidance for evaluators when making relation weightings can be put thus. First, some explanation of the alternative explanation framework would be in order. The guidance should then illustrate how stability of confidence is altered in line with this framework, possibly using the example of a formal model. As I have set out the framework in §5.4.1, I will not repeat it here. I have also provided an example of how one might go about adjusting one's stability of confidence (§7.3.1). The final thing to add would then be some questions for evaluators to ask when making a relation weighting. When evidence does not determine a precise degree of confidence, one must consider the instability of confidence. One should ask the questions:

Q: Would the criteria rule in an alternative explanation?

Result: If so, then the stability of confidence is decreased. If there is evidence of physical chance available, then one can calibrate the extent of stability of confidence to it. Otherwise, one uses expert judgement.

Q: Would the criteria rule out an alternative explanation?

Result: If so, then the stability of confidence is increased. If there is evidence of physical chance available, then one can calibrate the extent of stability of confidence to it. Otherwise, one uses expert judgement.

In my examples in the previous chapter, this reasoning process was followed when showing how the 'system dissimilarity' between MERS in rhesus macaques and humans results in a decrease in stability. An alternative explanation of the results was worked out, namely, the decrease in viral load observed was a result of the transient nature of the disease. I then explicitly considered some evidence of physical chance that may indicate the extent to which my confidence should become unstable. This was determined to warrant a 0.09 change in the negative direction only.

8.3.2 Chance

The balance of evidence was argued to determine what it is reasonable to believe. In qualitative terms, it was taken to mean how strongly one should believe in the hypothesis. I then argued that this should be constrained as far as possible to evidence of physical chance. The justification for using evidence of physical chance to determine degrees of confidence carries over from the justification for a calibration norm (§7.2.2). A calibration norm such as Williamson (2010)'s provides a method for calibrating degrees of belief to evidence of physical chance. In cases where a full application may be difficult, a more simple direct inference rule may be more appropriate. Which rule to apply will depend on context, but in all cases degrees of confidence should be assigned by the evidence of physical chance that is available. Instructions for making relation weightings falls out of this. Evaluators should, in the first instance, assign degrees of confidence based on evidence of physical chance. So degrees of confidence will be effected by criteria that are relevant to finding evidence of physical chance.

It is not so straightforward as to ask evaluators to consider whether a criterion provides evidence of physical chance. This is because there are multiple ways in which this can be done. Thus far, I have considered frequency data as the main candidate. I also noted that Williamson (2010)'s calibration norm allows calibration to more complex evidence of physical chance. In a similar vein, I claim here that there is another way in which degrees of confidence can be assigned, namely, through explanations. In Chapter 5, I appealed to McCain and Poston (2014)'s x-sphere example to justify the ability of explanatory content to affect one aspect of an agent's probabilistic profile, namely, the stability of probabilities/confidence. In this example, one agent's explanatory information meant that she did not alter her degree of confidence (0.5) that the next x-sphere would be blue, even in the light of a draw of another 10 blue x-spheres. The explanation of this is that her explanatory information renders her confidence stable. However, I will show here that there is a variant of the example that supports the idea that explanatory content can contribute to assigning degrees of confidence as well. My *x-sphere variant* is as follows.

Suppose that the physical set-up is exactly the same: an urn filled with red and blue x-spheres, that are drawn without replacement. Suppose further, at first, 15 x-spheres are drawn, consisting of 10 blue (B) and 5 red (R) x-spheres. This would lead Tom to assign $\mathrm{P}(B) = 0.66$. Sally would again not shift from $\mathrm{P}(B)$=0.5, as she has information that x-spheres have to always be stored in equal numbers of red and blue. In McCain and Poston's version, because of her information, Sally did not shift from the observed frequency data.

This shows how an alternative explanation contributes to stability of probabilities: her information offers an alternative explanation to the one that explains the observed results in terms of the proportion of x-spheres in the urn. In my variant however, Sally never shifts from $\mathrm{P}(B) = 0.5$ in light of any frequency data. It is not clear that her never shifting from 0.5 is because of the weight of evidence. It is true that her probability is less volatile. But it seems to be less volatile in an uninteresting sense. Uninteresting because the fact it never shifts from 0.5 implies that the probability was set at 0.5 by the explanatory information alone. McCain and Poston's argument was that frequency data sets probabilities, and explanations determine how resilient those probabilities are. But it seems that really, the explanation was what set the probability at the start. It seemed like explanations contribute only to stability in their example because the frequency data was in accord with the objective chance after 10 draws. But it is very likely that the frequency data would not be in accord with the objective chance after 10 draws, and in those cases, as my variant shows, it is the explanation that assigns probabilities.

On this count, the alternative explanations framework can be read as assigning degrees of confidence as well. One might propose separate guidance for relation weightings on the basis of this function of explanatory content. I will do so below, but first there are two difficulties to acknowledge. One is that the explanatory content seems to be doing no more than providing evidence of physical chance. This would subsume any guidance on this within the guidance above about evidence of physical chance. The other difficulty is that it may look like a special case of the stability norm, i.e., when explanatory content determines a precise objective chance then it affects degree of confidence, but when it picks out a range of possible chance hypotheses it affects stability of confidence. But this kind of distinction would tie stability of confidence to evidence of physical chance, rather than allow it to be effected by reductions in informativeness. Both problems boil down to whether there is a substantive difference between the process of explanatory content picking out a degree of belief and the process of it affecting the stability of belief. What then is this difference?

While it is true that in my variant of the x-sphere example, the explanatory information is providing evidence of physical chance, it does so as indirect evidence. The information that ‘x-spheres have to be stored in equal numbers’ explains why any probability for B must equal 0.5. In contrast, more direct evidence for B would be the frequency of blue balls drawn from the urn relative to the total number of balls drawn. The substantive difference between the two kinds of evidence is the reasoning processes involved in assigning probabilities. In the *direct evidence* case, one makes a direct inference, assigning probability 0.5 to B. In the *indirect evidence* case, the explanation is used to rule out alternative chance hypotheses other than $\mathrm{P}(B)$=0.5. This reasoning process will likely involve defeating other evidence of chance, namely frequency data that indicates a deviation from $\mathrm{P}(B)$=0.5. So if an evaluator is to use explanatory information to assign degrees of confidence, then they will do so by using a different reasoning process than one that involves using direct evidence of physical chance, and this must be accounted for.

However, the process of ruling in and ruling out alternative chance hypotheses to assign degrees of confidence seems to be identical to the process used to work out the extent of stability of confidence. This leads to the second problem: it seems like the role of explanatory content differs only in the element of confidence it affects, rather than differing by the reasoning process used, as I claimed. One might think that when explanatory information determines a precise value for the objective chance of an hypothesis it assigns a degree of confidence, but when it is consistent with a range of values then it assigns that range. We then interpret that range as the extent of the stability of confidence. This is the kind of idea that is at the heart of the imprecise probability model. Why then have separate instructions?

I think there is an important difference between the two reasoning processes. The basic idea is that stability of confidence is just a measurement of our uncertainty in what degrees of confidence to hold. But the reasoning process that affects stability of confidence does not select a range of values as *actual* degrees of confidence to hold. Instead it is a range of *possible* degrees of confidence. An example of a reasoning process that selects actual degrees of confidence is one that involves calibrating to evidence of physical chance. By contrast, a loss of informativeness does not pick out any actual degrees of belief to hold. This was the motivation for holding separate degrees and stability of confidence, as outlined in Chapter 5. By contrast, the process of ruling in an alternative chance hypothesis does involve selecting actual degrees of confidence. So the process of assessing one's stability of confidence is different to the process of selecting a belief state on the imprecise model. One does not just spread degrees of belief over probabilities consistent with the evidence of physical chance. Because one is assessing the relative informativeness of the evidence, the range of degrees of confidence 'opened up' will remain as possible degrees of confidence. This idea can be illustrated by another variant of the x-sphere example.

The set up remains the same. Suppose now that Tom, having set $\mathrm{P}(B) = 0.66$, finds out that the urn is filled with one kind of x-sphere, but does not know which kind. There are other kinds possible, including the kind that explodes if not in a 50/50 proportion of red and blue, as well as ones that do not explode, ones that just have to have more blues than reds, or ones that must have 80% blues. He now has less information than is needed to warrant a firm belief that $\mathrm{P}(B)$=0.66. But the explanatory information he does have leads him to rule in all chance hypotheses in the interval [0.5,0.8]. The only direct evidence of physical chance he has is $\mathrm{P}(B)$=0.66, so he sets his degree of belief to 0.66. As he has ruled in all chance hypotheses in [0.5,0.8] as *possible*, his degree of belief is unstable. Moreover, the extent of the instability is modelled by the interval [0.5,0.8]. Now, if he had the explanatory information that Sally had, he would set $\mathrm{P}(B)$=0.5. But his evidence is less informative than hers, so it is only compatible with the range. If one tried to provide only one set of instructions for going from explanatory information to both degrees and stability of confidence, one would ignore the differences involved in reasoning to actual and possible degrees of confidence.

Guidance for evaluators when making relation weightings between criteria and degree of confidence can be put thus. One should ask the questions:

Q: Would the criterion indicate evidence of physical chance?

Result: Degrees of confidence should be calibrated to this evidence of physical chance.

Q: Would the criterion provide an explanation for why current evidence of physical chance from other sources should not be calibrated to *and* provides direct information about the physical chance?

Result: Degrees of confidence should be calibrated to the explanatory information.

Q: Would the criterion provide an explanation for why the evidence of physical chance from other sources should be calibrated to, by providing direct information about the physical chance?

Result: Degrees of confidence should be calibrated to the evidence of physical chance from other sources.

8.4 Improving on EEMM

One might think the instruction outlined in the previous section is rather simple, and there is little need, beyond pedantry, to include it in the evaluation process guidance. This objection would have it that evaluators would be implicitly following this guidance anyway, and it would not take much special meta-methodological knowledge to do so. I respond to this kind of objection in this section, and show how the additions can improve on standard EEMM.

To begin with, although the questions seem rather simple they are arrived at by an extended analysis, carried out over three chapters of this thesis. This analysis considered many disparate areas of philosophy and scientific practice: the distinction between weight and balance of evidence; the implications of this distinction for evidence evaluation in medicine; a number of competing varieties of Bayesianism; how formal models can be applied to evidence evaluation; the suitability of formal models for evidence evaluation; the reduction of subjective choice in evidence evaluation. While the questions may be simple, the theoretical motivations are not.

One might still think that there is no benefit to adding these questions. In the main, evaluations will proceed broadly correctly. Justification for this view could be obtained by considering the links between the interpretations of *quality* and *status* and the status criteria. For example, lowering status due to considerations of a future evidence base should lead to a change in the stability of confidence. Reasoning about future evidence bases requires reasoning about alternative evidence bases. It is not too much of a leap to think evaluators would be implicitly reasoning about alternative explanations. However, it is not so clear that evaluators following the standard EEMM process would assign degrees of confidence due to following a reasoning process that is similar to the one I have outlined in this chapter. To develop this idea, I will compare the judgements I made on the status of the mechanism claim from my systematic review in Chapter 3, with judgements that would be made if the instructions for making relation weightings developed in this chapter were followed. The comparison of judgements will show why the additions are useful because the systematic review was carried out using only the guidance found in EEMM, plus the kind of expert knowledge that is likely to be held by an evaluator in the relevant field. The original review was carried out in this way in order to highlight how the additions will improve EEMM.

At the start of this chapter, I used the status criterion ‘there are many sources of evidence for crucial features’ to illustrate the idea of a relation weighting. There, I posited relation weightings between this criterion and both degree and stability of confidence. In my review, I evaluated my degrees of confidence in the mechanism claim by appealing to this criterion. However, it does not meet either of the questions that would determine a relation between a criterion and assigning degrees of confidence. This is one discrepancy between my review and one that utilises the questions formulated in this chapter. For sure, qualitatively, meeting this criterion means the evidence is *for* the claim. The question is whether it makes the evidence stronger on account of assigning high degrees of confidence, or on account of stabilising confidence. On the relation weighting guidance, and the reasoning principles that support them, the information that there are 'many sources of evidence' should not directly affect one's degrees of confidence. Answering 'yes' to this criterion would not directly provide evidence of physical chance, nor would it provide extra explanatory information about the physical chance. The fact that we have evidence for 'crucial features' may be more likely to provide evidence of physical chance than the fact there are many sources. This might be because we can use the evidence of physical chance for that crucial feature to stand in for the evidence of physical chance for the mechanism claim. However, this would require explicitly setting out in the guidance that it is not the mere fact that we have evidence for crucial features that should result in holding high confidence in the claim. Instead the guidance should state that it is the evidence of physical chance that matters.

This is one example of where my review did not make the right relation weighting. There is another aspect to this discrepancy. The 'there are many sources of evidence for crucial features' criterion may also affect stability of confidence. At first sight, the phrase 'multiple sources' might look like a way of saying this evidence has a high gross informativeness. However, as argued in the previous section, the relevance of the amount of studies in an evidence base to its informativeness is in the contribution to net-informativeness. It is unlikely that an evidence base will consist of so many sources of evidence that it is maximally informative. So, an evidence base with many sources will have higher net informativeness than one with only a few sources of evidence. Plausibly, if all studies point in the same direction, then the width of the probability interval will be narrower than on any one study alone. There is therefore another level to the discrepancy between how

I utilised this criterion, and how it should be used. I should not have used it to assign degrees of confidence, and I should have levelled it as support for increasing the stability of confidence.

A natural question to ask in the light of this discrepancy is: would the addition of the relation weighting guidance change the review and improve it? A case might be made that the evidence of physical chance does point strongly in favour of the claim, so the two judgements would not differ. In my running example one option for the degree of confidence was 0.84. However, a competing option was 0.51 through the application of the Equivocation norm. I argued that 0.51 was arguably better motivated. Additionally, one might think that 0.84 is not high confidence, but rather moderate-to-high. So on either of the values, the review that implements my additions would get a lower degree of confidence than one that did not. This difference alone is not a reason to prefer a guidance that has my additions. Part of the reason why one would prefer the additions is no different to the basic motivations for calibration and non-evidential norms (see §§7.2.2 and 7.3.1). The norms that stipulate how evidence constrains belief find their main motivation in the task of restricting subjectivity. So the additions are motivated with principles that are important for reducing subjectivity in evidence evaluation. Expert judgement will not always follow these norms. Additionally, the additions are motivated by considerations pertaining to the weight and balance of evidence. These aspects of evidence are important to acknowledge because they represent two different dimensions of the strength of evidence. And only on my additions can we be sure that an evaluation would capture both dimensions. Therefore, an evaluation that is consistent with the original guidance would be at more risk of subjective choice, as one can choose to use the information that one has 'many sources of evidence' to assign degrees of confidence. What counts as 'many' may be highly dependent on the expert's background knowledge. It was fortunate that in my review, high(ish) degrees of confidence are still compatible with the evidence. But it is easy to imagine a case where there are many sources of evidence in favour of a claim, an evaluator thus holds high confidence in the claim, but the evidence of physical chance itself only supported holding moderate degrees of confidence. This would be the case if a probabilistic analysis found that the evidence only made the hypothesis moderately probable. In such a case, evaluator subjective choice results in a stronger judgement than otherwise should have been made.

While following the guidance led to a mistake, my review also got some relation weightings for status criteria right. When evaluating the status of the mechanism claim, I also considered whether use of analogy would lower status. One of the major issues was the use of a rhesus macaque model, which is inferior to a common marmoset model of MERS infection. I considered this to keep quality low, because if a common marmoset model was used in future, our confidence in the claim is likely to change. This seems like a correct relation weighting to make, given the questions that are about alternative explanations. That the rhesus macaque model capitulates only mild and transient MERS is relevant to the claim that a mechanism exists between combination therapy and recovery, because those patients who die typically have severe MERS. Indeed, in the running example of §7.3.1, I used this feature of the evidence to demonstrate how stability of confidence would be modelled.

Consideration of how I got this relation weighting right does however lead to a potential problem with my analysis. One might think I got this weighting right because it is easy to implicitly make relation weightings, and the example of discordance above was just a simple error. However, it seems more plausible that I got this weighting right due to it being the case that none of the status criteria affect degree of confidence. No criterion explicitly asks evaluators to consider the evidence of physical chance available. Instead, they only seem to affect stability of confidence. A consequence of this is that an evaluation by status criteria would really only be lowering the quality of evidence aspect of the status interpretation. This would lead to a change in the process where the status criteria are raising or lowering status only by means of affecting the quality of evidence, and so stability of confidence. This would also mean that the entire evaluative process found in EEMM only really considers the quality of the evidence. The degree of confidence one holds in the claim is evaluated by another means, namely, by criteria that would fulfil the two questions about evidence of physical chance from the previous section.

This would be quite a strong change to the process. Moreover, it might make the process so restrictive as to make an evaluation less reliable. Recall that any additional guidance must not be overly restrictive such that an evaluation would suffer as a result. Arguably, by making none of the evaluative criteria relevant to the degree of confidence, the evaluation would be diminished. One reason to think this, is that it excludes expert judgement and makes the process of assigning degrees of confidence a purely statistical question. This might lead to accusations of arbitrariness when our evidence is not so specific that we can justify assigning degrees of confidence on evidence of physical chance alone.

To counter this issue, I suggest a compromise. This is in a similar vein to the compromise I suggested in the previous chapter, where problems with applying formal models led to the suggestion of approximating idealised models. My proposal is that within a formal EEMM there is still scope to use judgement to assign degrees of confidence. A slight alteration to the additions I propose in this chapter might be that one should first assign degrees of confidence in accord with evidence of physical chance first. Then, if the status criteria indicate that one's degree of confidence should be greater than is assigned by evidence of physical chance, then one should select a further more precise interval to select degrees of confidence from. For illustration, consider the running example from the previous chapter. Evidence constrains one's degrees of confidence to $I = [0.51, 0.95]$. But the fact that there is a lot of evidence for crucial features would mean restricting this interval further in the next stage of the process. One might think that $b \in [0.7, 0.8]$. One would then select degrees of confidence using the preferred non-evidential norm. If it were Equivocation, then $b = 0.7$. So this process would result in a higher degree of confidence than if one did not follow this process (0.51), which seems to accord with how the status criteria are supposed to work in the standard process. Moreover, it keeps in line with my proposal, as one cannot choose degrees of confidence that fall outside of what was first compatible with the evidence of physical chance.

This would of course need a fuller development. Any different proposal would need to show that subjective choice is decreased to a greater relative extent than in the one developed here. But the take home lessons are: the status criteria primarily affect stability of confidence; an evaluation that puts more emphasis on assigning degrees of confidence through calibration to evidence of physical chance would be better; there is still room to use the criteria to assign degrees of confidence. This keeps in the spirit of the semi-formal approach to evidence evaluation that I advocate.

8.5 Conclusion

In this chapter I identified part of the EEMM process where subjective choice may enter, namely, at the making of relation weightings. To mitigate this problem, I formulated some additional guidance on relation weightings. I used the norms that govern how evidence imposes constraints on belief from Chapter 7, and the features of evidence that contribute to the weight and balance of evidence from Chapter 5, to formulate two sets of questions. One set guides evaluators to look for criteria that assign degrees of confidence through consideration of evidence of physical chance. The other set guides evaluators to look for criteria that contribute to the stability of confidence through consideration of alternative explanations. I then argued that my additions would improve on EEMM, by comparing some judgements from my systematic review from Chapter 3, with judgements using the same criteria but with my additions. I claim that my additions allow subjective choice to be limited to a greater extent than on the original guidance.

Chapter 9

Integration with GRADE

9.1 Introduction

Thus far, I have focused mainly on the evaluation of mechanism claims. EBM+ is however a methodology for causal evaluation and EEMM is ultimately concerned with evaluating causal claims. This involves combining statuses of mechanism and correlation claims (see §§3.2 and 3.4.4). Combining the statuses of claims may at first seem fairly straightforward. One is instructed to take the minimum of the statuses of the correlation and mechanism claims to be the status of the causal claim. However, one must first evaluate the correlation claim, and EEMM exports that task to extant evaluative frameworks that focus on clinical studies.

In this chapter I consider how to extend the EEMM approach through combining it with an evaluation of clinical studies on the GRADE framework. I have used GRADE as a case study throughout this thesis and in §9.2 I argue that it is a viable candidate for evaluating clinical studies. In doing so, I respond to some independent critiques of GRADE

(§9.2.1). GRADE cannot alone evaluate a correlation claim, as both evidence from clinical and mechanistic studies can provide evidence of correlation. Thus, in §9.2.2 I argue that the guidance for evaluating the status of a correlation claim on both clinical and mechanistic studies is incomplete. I suggest ways to use both EEMM and GRADE to improve on this. In §9.3 I identify a potential compatibility issue between EEMM and GRADE, namely the use of effect size information. On the basis that effect size information is useful for decision making and for directing future research, I argue that effect size information should be included when evaluating the status of a causal claim on EEMM (§9.3.1). This leads to considering how to adjust effect size information in response to finding quality issues with the evidence (§9.3.2). I argue that EEMM should be extended by including the *certainty range* approach, which was recently proposed by GRADE as a way to probabilistically represent confidence in interval estimates. I address some potential conceptual problems with this approach, and indicate how one may implement it in practice. How to fully operationalise this approach has not been settled on even by GRADE, so this area presents a fruitful avenue to explore for future research on EBM+, and evidence evaluation broadly construed.

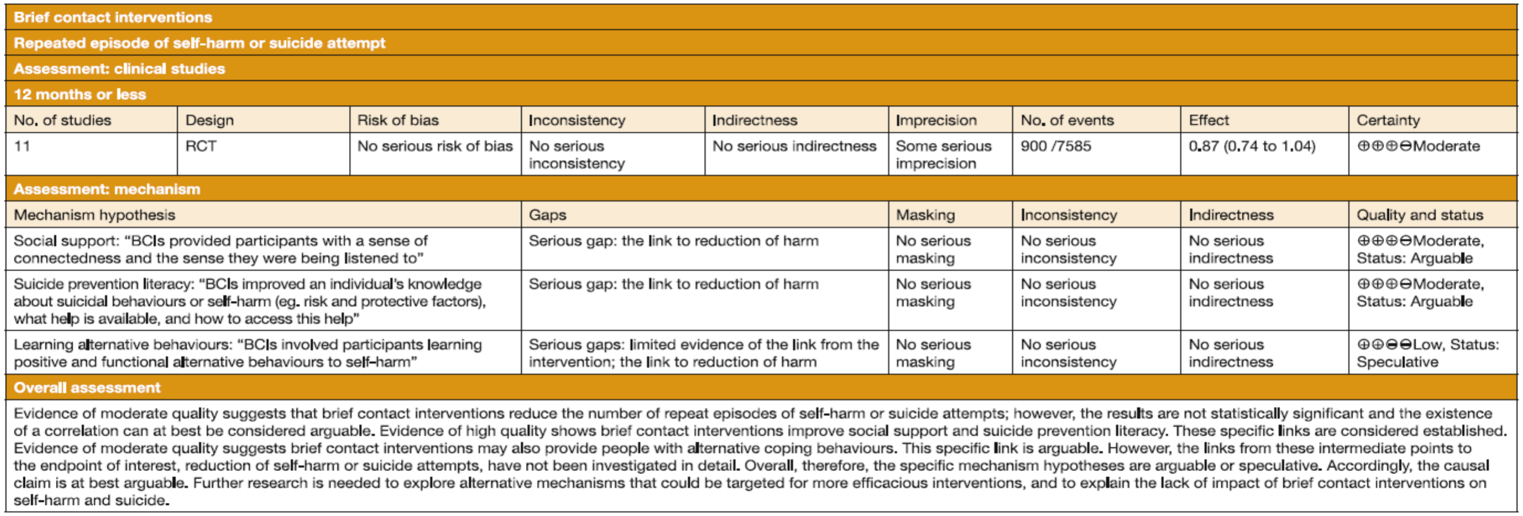
9.2 GRADE: Evaluating correlation claims

In this section I look at whether GRADE can be taken as a way to evaluate correlation claims such that evaluations carried out on the framework can be integrated into EEMM. I discuss some reasons for using GRADE, and then address some criticisms of GRADE as an evidence evaluation framework. I then argue that GRADE can be seen as a clinical trial evaluator. However, to evaluate a correlation claim, one needs to include evidence from both clinical and mechanistic studies. I identify that guidance on this is lacking in EEMM and sketch a way to improve on this part of the process. Full development of these ideas is likely to benefit from collaboration between those working on the methodology of clinical and mechanistic study evaluation.

9.2.1 Why GRADE?

There are a number of reasons why you would choose GRADE to evaluate a correlation claim. Firstly, the structure of the evaluation process on GRADE and EEMM is similar. In both frameworks, one uses evaluative criteria to identify inadequacies or positives in the evidence base, a process which leads to an evaluative judgement put in terms of one's confidence in some claim. I have explicitly demonstrated this consistency in Chapter 5, albeit with a slight alteration to GRADE, namely, that its outputs should be in terms of both degree and stability of confidence. This means that when evaluating a correlation claim using GRADE, one would not be following a reasoning process completely different to the one used to evaluate a mechanism claim. Relatedly, as identified in Chapter 5, GRADE is compatible with the evaluative judgements used in EEMM. Judging one's stability and degree of confidence is possible on GRADE, and as I argued in that chapter, should in fact be the way GRADE characterise their judgements on the strength of evidence. Finally, EEMM provide GRADE style tables (see fig. 9.1). This is a standard way that GRADE require users to summarise evidence. The fact that EEMM can also be put in this format provides further evidence that the two frameworks are compatible.

Unfortunately, there are a number of criticisms of GRADE other than the ones I have dealt with in this thesis (see §§5.5 and 6.4). Firstly, GRADE consider their evaluation framework to be a causal evaluation framework, in line with standard EBM. So the outputs are supposed to be about a causal, not a correlation, claim. But they only explicitly consider evidence obtained from clinical studies. This runs counter to what EBM+ require, and is a stance on evidence that I have rejected in this thesis. In line with evidential pluralism, defended in this thesis and elsewhere, GRADE cannot be a causal evaluation framework as it does not include evidence from mechanistic studies. Mercuri et al. (2018) also raise this problem, and it can be avoided if we see GRADE as being restricted to assessing only evidence from clinical studies. It is no problem to restrict quality assessment frameworks to one kind of evidence. Indeed, the majority of the guidance in EEMM is on



Brief contact interventions								
Repeated episode of self-harm or suicide attempt								
Assessment: clinical studies								
12 months or less								
No. of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	No. of events	Effect	Certainty
11	RCT	No serious risk of bias	No serious inconsistency	No serious indirectness	Some serious imprecision	900 /7585	0.87 (0.74 to 1.04)	⊕⊕⊕⊖Moderate
Assessment: mechanism								
Mechanism hypothesis			Gaps		Masking	Inconsistency	Indirectness	Quality and status
Social support: “BCIs provided participants with a sense of connectedness and the sense they were being listened to”			Serious gap: the link to reduction of harm		No serious masking	No serious inconsistency	No serious indirectness	⊕⊕⊕⊖Moderate, Status: Arguable
Suicide prevention literacy: “BCIs improved an individual’s knowledge about suicidal behaviours or self-harm (eg. risk and protective factors), what help is available, and how to access this help”			Serious gap: the link to reduction of harm		No serious masking	No serious inconsistency	No serious indirectness	⊕⊕⊕⊖Moderate, Status: Arguable
Learning alternative behaviours: “BCIs involved participants learning positive and functional alternative behaviours to self-harm”			Serious gaps: limited evidence of the link from the intervention; the link to reduction of harm		No serious masking	No serious inconsistency	No serious indirectness	⊕⊕⊖⊖Low, Status: Speculative
Overall assessment								
Evidence of moderate quality suggests that brief contact interventions reduce the number of repeat episodes of self-harm or suicide attempts; however, the results are not statistically significant and the existence of a correlation can at best be considered arguable. Evidence of high quality shows brief contact interventions improve social support and suicide prevention literacy. These specific links are considered established. Evidence of moderate quality suggests brief contact interventions may also provide people with alternative coping behaviours. This specific link is arguable. However, the links from these intermediate points to the endpoint of interest, reduction of self-harm or suicide attempts, have not been investigated in detail. Overall, therefore, the specific mechanism hypotheses are arguable or speculative. Accordingly, the causal claim is at best arguable. Further research is needed to explore alternative mechanisms that could be targeted for more efficacious interventions, and to explain the lack of impact of brief contact interventions on self-harm and suicide.								

Figure 9.1: GRADE style assessment table from Parkkinen et al. (2018b, p.55), completed for an assessment of brief contact interventions for reducing self harm.

evaluating evidence from mechanistic studies. So this alteration to seeing GRADE as a clinical study evaluator is also not a problem.

Another specific criticism of GRADE is that there is no justification for the concepts it uses or the evaluative processes central to the framework. For example, Mercuri and Gafni (2018a,b,c) argue that there is no justification for which moderating domains should be part of the process (details of domains can be found in Table 1.1 on page 22). Neither is there justification for how assessment on those domains results in specific extents of quality reductions, e.g., why should a serious breach of randomisation result in a decrease of one quality level? However, GRADE's evaluative criteria were identified "because they address nearly all issues that bear on the quality of evidence [for clinical studies]" (Balshem et al., 2011, p.405). So the justification for this part of GRADE derives from broader clinical research principles. As noted in §4.3, there is a wealth of standardised information on the methods used in clinical trials, exemplified by the existence of textbooks on the subject (e.g., Fletcher et al. (2012)). Mercuri and Gafni (2018a,b,c) would have to argue that one cannot draw on this wealth of information to justify the GRADE process. However, the domains were:

> "arrived at through a case-based process by members of GRADE, who identified a broad range of issues and factors related to the assessment of the quality of studies. All potential factors were considered, and through an iterative process of discussion and review, concerns were scrutinized and solutions narrowed by consensus to these five categories." (Balshem et al., 2011, p.405)

Reliance on expert judgement is also part of the justification for the extent of effect different moderating domains may have on quality ratings. So Mercuri and Gafni may object to this process, which relies on the background knowledge of the GRADE working group. However, as I have argued elsewhere in this thesis, expert judgement is integral to evidence evaluation, but can be constrained. Mercuri and Gafni would have to show why expert judgement results in too much influence by subjective choice. It is not enough to just identify that expert judgement is playing a role in developing the framework. So GRADE does have justification for its criteria *and* how those criteria impact ratings, namely the expert knowledge of the GRADE working group and the evaluators who use GRADE, respectively.

9.2.2 Evaluating correlation claims

I have argued that GRADE can be seen as a clinical trial evaluator. It can then be used to arrive at a status of a correlation claim. However, EEMM exports only *most* of the evaluation of the correlation claim to extant evaluative frameworks. Mechanistic studies can also provide evidence of correlation, so part of the evaluation of a correlation claim can be carried out on EEMM. It is also the case that causation can be established on the basis of clinical studies alone (§2.2.3). So to be precise, GRADE is an 'evidence of causation on the basis of clinical studies' evaluator. When the evidence from clinical studies is sufficient to establish both a mechanism and correlation claim, then it also establishes causation. When it is not sufficient to do so, then GRADE can be used to output a status for the 'correlation claim on the basis of evidence from clinical studies'. The 'correlation claim' must be evaluated by both evidence from clinical and mechanistic studies. This poses a problem for EEMM, as it does not provide much systematic guidance on how mechanistic studies can be used to evaluate the status of a correlation claim. As identified in §4.2, there is some guidance for how mechanistic studies may 'boost' the status of a correlation claim. But for GRADE to be able to play a role in evaluating a correlation claim, one might think that there is a need for status criteria for such claims. This is also the case for criteria for evaluating a correlation claim on the basis of clinical studies. Note that GRADE could also be used to output a status for the 'mechanism claim on the basis of evidence from clinical studies'. As the guidance for evaluating a mechanism claim is much more complete, I will focus below on the problem of providing status criteria for correlation claims.

There is a short section in EEMM that indicates how one should evaluate the status of a correlation claim. Parkkinen et al. (2018b, p.92) states that once one has identified a set of plausible confounders and carried out "an assessment of the quality of the design of the relevant studies, deciding whether the putative cause and effect are correlated is a purely statistical question". For example, when a meta-analysis of the studies in an evidence base has been carried out, the status of the claim will depend on "the width of the confidence interval, the size of the p-value, and the heterogeneity of the studies evaluated" (ibid.). This looks like a set of status criteria, although it is not very systematic. But the focus on statistical measures in determining status is consistent with the idea that degrees of confidence in a claim should be set by evidence of physical chance. However, there is more to a GRADE evaluation than this. There is of course a quality assessment. And this assessment goes beyond the 'design of relevant studies', as it takes into account all factors that matter for determining the net informativeness of the evidence base. So a fuller account of what it takes to evaluate a correlation claim would have to include all these features.

There are other reasons to think that there is more to evaluating a correlation claim than the purely statistical question. Firstly, not all evidence bases will have undergone a meta-analysis. Indeed, GRADE evaluates only evidence from clinical trials, rather than statistical aggregation methods such as meta-analyses. Secondly, it is plausible that there will be questions to ask about the correlation claim itself, rather than just the quality of the studies under evaluation. In the case of establishing a mechanism, one asks questions about what the evidence says about the mechanism itself and not just about the studies. For example, one considers whether the mechanism exhibits a high degree of complexity. Arguably, this will also be the case for correlation claims. Indeed, Williamson (2019) argues that it is not enough to just establish that two variables are probabilistically dependent in the sample of outcomes in a trial to establish a correlation. One must work out whether it is a genuine correlation in the underlying data-generating distribution. This could be one kind of status criterion for a correlation claim. Now, Williamson notes that this may depend on the method of sampling and size of sample. And these are issues that may be picked up on and resolved in a meta-analysis. But as already stated, a meta-analysis may not always have been conducted. Moreover, meta-analyses are not themselves without potential quality issues (Stegenga, 2018; Holman, 2018).

For the purpose of adding status criteria for correlation claims to EEMM, one option is to make specific reference to the evaluative criteria of the GRADE framework. There are two problems here. One, GRADE does not assess the quality of meta-analyses, so solving this issue is not as simple as deferring to the GRADE process. It will likely require more information on what it takes to evaluate the status of a correlation claim, and additionally whether a separate evaluation of statistical aggregation methods must be carried out. Two, this option may limit applicability to areas that GRADE currently do not account for. The framework is designed with intervention assessment in mind, rather than, for example, exposure assessment. What would be helpful is to make reference to the fact that one needs to export the majority of an evaluation of the quality of evidence from clinical studies to an extant framework. And then provide systematic guidance on what sort of criteria contribute to raising or lowering the status of a correlation claim, beyond what can be said about assessing individual studies and the 'purely statistical question'. Providing concrete examples of such criteria is beyond the scope of this thesis, which is primarily concerned with the evaluation of evidence from mechanistic studies. Additions in this area would however be helpful, and this kind of task is likely best exported to those already working on the evaluation of clinical studies. Collaboration on this task with those working on evaluation of evidence from mechanistic studies is likely to be needed, as a conceptual shift is required from evaluating evidence in order to establish causality to evaluating evidence in order to establish a correlation.

Another alteration to GRADE is in order with respect to assigning statuses to correlation claims. Recall from Chapter 5 that this judgement should reflect both the likelihood that one's confidence in a claim should change and the extent of that confidence. GRADE already has a method for making a judgement on the stability of confidence, as I have argued that the quality assessment process should do this. To adapt the GRADE framework one would also need some way of assigning a degree of confidence to the claim, which at the same time is not part of the moderating domains that affect stability of confidence. Originally, this was done on the basis of study design. I responded in §5.5 to a criticism made by Stegenga (2018) and Mercuri et al. (2018) of how GRADE rates quality first on study design and then on moderating domains. I have suggested that GRADE remove this initial part of the process and start evidence from all kinds of studies at the same degree of confidence. I also argued in Chapters 7 and 8 that degrees of confidence should, where possible, be calibrated to physical evidence. So a simple method for assigning degrees of confidence on GRADE would be to use such evidence, e.g., p-values, frequency data, etc. For most clinical studies, the statistical significance threshold for p-values is $p \leq 0.05$. As the studies that are considered good enough to be evaluated will already have been found statistically significant, calibrating b to $(1-p)$ will result in $b \geq 0.95$. However, this may not always be the case, as there is also the option for explanatory information to help set degrees of confidence, which may result in assigning a lower degree of confidence. So, to make my alteration to GRADE more precise, degrees of confidence are first set by evidence of physical chance or explanatory information, rather than by the kind of method used in the study. This would mean that evidence from all kinds of studies will not always start at the same degree of confidence, but it will be the evidence of physical chance, rather than the study design, that sets it.

In sum, GRADE can be used to output a status for a correlation claim by considering how strongly the evidence of physical chance is in favour of the claim and calibrating one's degree of confidence to it. The standard moderating domains can then be used to work out how stable one's confidence is. Other *status criteria* for correlation claims may also be possible, but full formulation of them is outside of the scope of this thesis.

9.3 Effect sizes and extensions

GRADE can be used to evaluate clinical studies, and with some addition to EEMM can be combined with evaluation of evidence from mechanistic studies to assign a status to a correlation claim. In this section I consider a specific set of worries that arises from consideration of the role of effect size information in the GRADE process. I first argue that effect size information should be included in an evaluation of all claims on EEMM. I will then consider how EEMM can be extended to both include and adjust effect sizes as part of an evidence evaluation.

9.3.1 Motivation for including effect sizes

Effect sizes are central to a GRADE evaluation: one rates one's confidence in effect estimates. There is however some inconsistency with how EEMM uses effect sizes. The mechanism claim is specifically characterised as

> "there is a mechanism linking the putative cause A to the putative effect B that explains instances of B in terms of instances of A, and which can account for the observed correlation between A and B." (Parkkinen et al., 2018b, p.16)

Elsewhere mechanisms must also “account for the magnitude of the correlation” (Parkkinen et al., 2018b, p.15) or “account for the extent of the correlation” (Parkkinen et al., 2018b, p.28). At first sight, the mechanism claim need only account for the existence of a correlation, but in the text that explains the evaluation process (which is found in the latter two quotes), what it takes to account for a correlation is to account for the existence of a correlation and the effect size. Complicating this issue is that the characterisation of what it is for two variables to be appropriately correlated does not make reference to the effect size (§1.1). Neither does the characterisation of a causal claim, e.g., an efficacy claim is defined in EEMM as a claim that an intervention or exposure has some specific effect. This formulation is unclear on whether this ‘specific effect’ is a specific magnitude, or just that a specific kind of effect exists (e.g., a decrease in blood pressure). It is also the case that evidential pluralism, the thesis that motivates EEMM, does not make reference to effect sizes. As identified above, the effect size is relevant for *establishing* a correlation claim, but it is not a part of the claim itself. When integrating a GRADE and mechanism claim evaluation, then, should the claim include information about the effect size? There are two reasons one might think that it should. One appeals to decision making, and the other involves the relative importance of different magnitudes of effect sizes. These reasons are interrelated and I will explain each in turn.

Thresholds are an important idea in decision making: belief thresholds that must be surpassed in order to make a decision were central to arguments made in §2.3.2 and §7.2.3. *Critical thresholds* are a concept that is unique to GRADE (Schünemann, 2016; Hultcrantz et al., 2017). They are not belief thresholds, but instead are an effect size threshold over which one can deem an effect beneficial. The basic idea is motivated by the fact that on GRADE, a decision on whether an intervention should be recommended is not made on the absolute strength of the evidence that would support it, but rather on the *adequacy* of the evidence to support that recommendation. Critical thresholds are how one determines adequacy. This threshold is assigned a value equal to the delta (δ) for an intervention, where δ is a pre-defined level at which an effect for a specific intervention is deemed

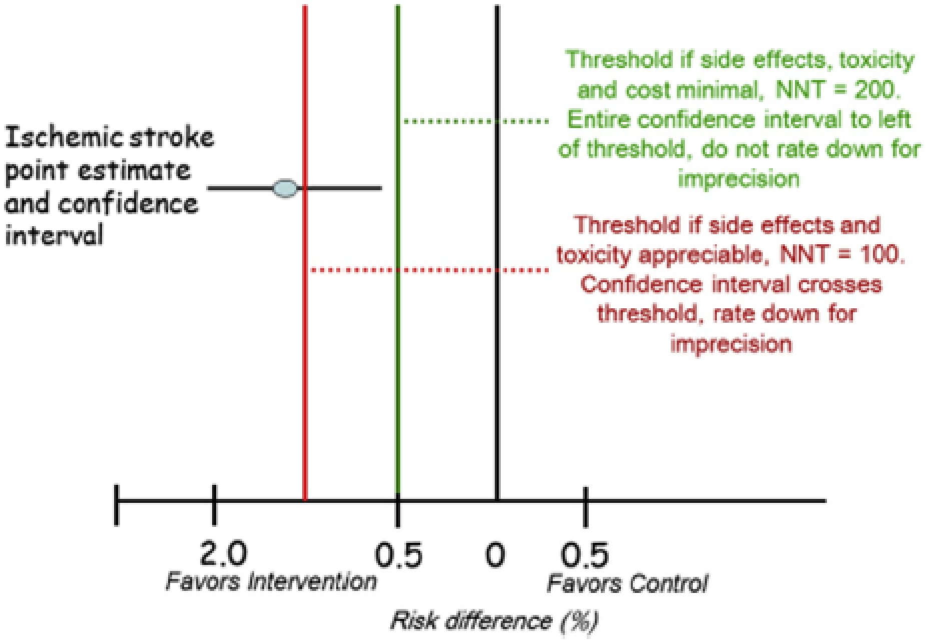

Threshold if side effects, toxicity and cost minimal, NNT = 200. Entire confidence interval to left of threshold, do not rate down for imprecision
Ischemic stroke point estimate and confidence interval
Threshold if side effects and toxicity appreciable, NNT = 100. Confidence interval crosses threshold, rate down for imprecision
2.0
0.5
0
0.5
Favors Intervention
Favors Control
Risk difference (%)

Figure 9.2: Visual illustration of critical threshold approach from Hultcrantz et al. (2017). The green and red lines are different values for δ, given different harm profiles. The confidence interval is wholly above the green line. In that context, this estimate would be adequate to base a recommendation. If the harm profile of the intervention was greater, represented by the red line, then the interval estimate would not be adequate to base a recommendation.

beneficial. This is worked out by consideration of the trade off between known harms and potential benefits of the intervention. As introduced in §5.4.2, GRADE's imprecision domain becomes especially relevant when δ is considered. An interval estimate may look very wide, but if it does not cross the δ for that intervention, then it does not do badly on the imprecision domain. This is because the true effect, on the assumption it is a value within the estimate, is still likely to be beneficial. In the context of decision making, evidence may be considered adequate to make a recommendation when the estimate is wholly above δ. This process is visually demonstrated in fig. 9.2.

The importance of critical thresholds transfers over to effect size information. If we did not include effect size information in our causal claim evaluation then we might miss out on making decisions that we should have made. For illustration, suppose that an interval estimate I for drug D is [1.5,2.5] where I is measured on an arbitrary scale that goes from 0 to 10. We work out that given the relatively low harm profile of D, its δ is 0.25. If we did not have δ, one might be tempted to say that this effect size is small, and on that basis be reluctant to recommend D. But given that we do have δ, we may be inclined to recommend this treatment. Equally, we might make recommendations when we should not have, if it were the case that what counted as a large or small effect size is assessed without tying it to a specific δ. Even large effect sizes may not lead to recommendation if the critical threshold is also quite large. So if the status of a causal claim is merely about the existence of a causal claim, then there could be cases where the effect size of an established causal claim is lower than the δ for that intervention. If we were to recommend such a treatment on the basis that causality is established, our recommendation would be misguided.

One might object that EEMM makes no recommendations for decision making. And once it has been established whether causation holds or not, it is up to a guideline panel to use effect size information to make a decision. However, central to EEMM and evidence evaluation more broadly is the requirement to make all judgements explicit. Additionally, EBM is centrally concerned with providing guidance on how to make decisions. If EBM+ is to be an improvement on EBM, then it needs to both make everything relevant to evaluating causation in medicine explicit and say something about how evidence integrates into decision making processes. Given that effect sizes are important for decision making purposes, they should be included in the EEMM process.

One may then question whether decisions should be made on evidence demonstrating a small effect size, even when it surpasses some critical threshold. Indeed, there are many criticisms of small effect sizes: we saw one in Chapter 6, where one of the arguments for the medical nihilism thesis was that small effect sizes are ubiquitous, and this has negative consequences for medicine.

There are however reasons to resist this wholesale critique of small effect sizes. One reason is that the magnitude of an effect should not matter if it is to have some net-benefit to a patient. As argued in §6.3, curing a disease is not the sole goal of an intervention. A small effect size for a clinical outcome might nonetheless improve the quality of life of a patient. If the effect size is some quality of life measure, then a small effect size may be an issue. But it is not clear in critiques of small effect sizes whether the problem is relative to the kind of outcome. For example, Stegenga (2018, p.173) notes that statins, which are preventative drugs for cardiovascular diseases, have been found to benefit only 1.2% of patients. He does not state whether this benefit is relevant to quality-of-life or specific reductions of cardiovascular-related morbidity. Even if it is a quality of life effect, 1.2% may still be an effect worth having, given that statins are widely prescribed, and 1.2% of patients will be a large absolute number of patients who benefit.

Another reason is that point estimates of effect sizes are statistical averages. It could entirely be the case that a small effect size means the intervention has a large effect for a subset of patients, and little to no effect for others. This is of course captured in the confidence interval approach, where the interval can be used to indicate the variance of measured effects. Wide intervals are often taken as indicating that there is something wrong with the evidence, e.g., that the sample size is too small. But this may not always be the case. Variability in effect may exist, independent of the size of the sample. And in neither the point or interval estimate case does this always mean that the intervention is ineffective. As we start to understand more about the variance of responses for individual patients to interventions, or the variance in disease states for sufferers of 'one' condition, these variances in effect size matter more. For example, acute myeloid leukemia (AML) is a single disease as it has a specific characterization: "blocked myeloid lineage differentiation and accumulation of leukemic blast cells" (Assi et al., 2019, p.151). But it is a 'highly heterogeneous disease' as it can be caused by many different types of genetic alterations. These different genetic alterations define separate sub-types of AML, and patients with different sub-types may require different specific treatments (Assi et al., 2019). Testing interventions that have shown promise for treating AML may only give small effect sizes as a result of not targeting all kinds of sub-types. But to conclude that such an intervention is not effective would be too hasty. The wide degree of variance observed in a study that tests such an intervention may just indicate that there is a sub-set of patients for whom the drug works. Therefore, to know that there is an effect at all is valuable. This is the case for decision making if the effect size is above some pre-defined δ, as there will likely be many other patients for whom the drug is much more effective. And it is also the case for directing research into which sub-type it may be effective for. This latter reason provides independent motivation for including effect sizes in EEMM. I argued in §4.2 that providing information to direct future research is a valuable part of an evidence evaluation process. That remains the case here. Hence, effect sizes should be included in evidence evaluation, including in EEMM.

9.3.2 How to include effect sizes

Certainty ranges

Having justified the inclusion of effect size information, it is next natural to question *how* this would be included in an evaluation. At present, the status judgement is formulated in terms of having confidence in some claim C, e.g., A causes B. When we shift to C_ε = 'A causes B with effect size ε', we might be tempted to just say we hold confidence that is approximate to $P(C_\varepsilon)$. Another approach has recently been proposed by GRADE. This involves a turn towards a *semi-quantitative* measure of one's confidence in effect estimates (Schünemann, 2016). This concept is not yet fully endorsed by the GRADE working group. However, it is under development and has been applied to a real systematic review (Tikkinen et al., 2017). Specifically, one's confidence is conceptualised as a *certainty range*, which is a quantitative measure of uncertainty spanning the interval estimate and altered by consideration of moderating domains (Schünemann, 2016; Tikkinen et al., 2018). Tikkinen et al. (2018) unpack the reasoning used in the application of the certainty range approach to a review of thromboprophylaxis interventions in urological surgery (Tikkinen et al., 2017). But the basic idea is that one starts with a probability distribution defined over the interval estimate, which will be derived from the statistical analysis performed in

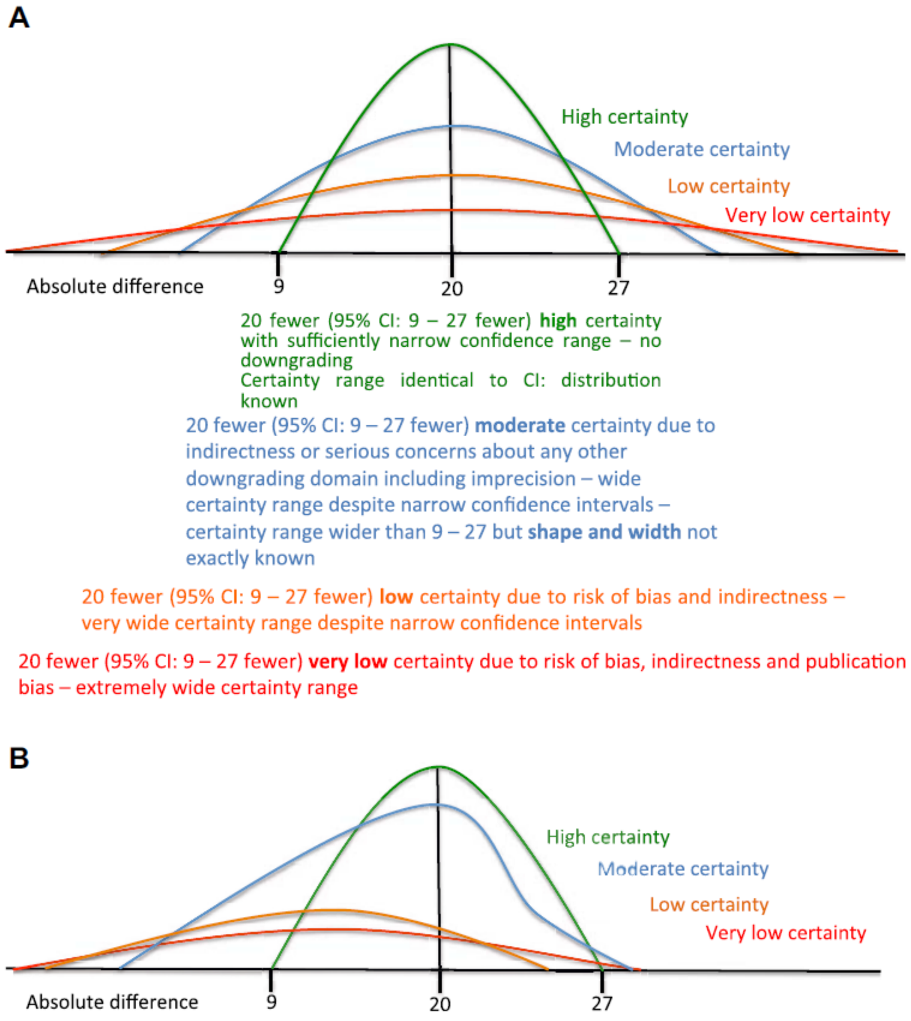


Figure 9.3: Visual illustration of the concept of certainty ranges from Schünemann (2016). Initially, the certainty range is assumed to be normally distributed over the interval. Quality issues then widen (**A**) or skew (**B**) the certainty range. On altered ranges, more probability is lumped over values outside of the interval estimate relative to values inside the estimate.

the study. The initial certainty range is represented by this probability distribution, and one then evaluates the evidence and adjusts the certainty range. Adjustments are made using evidence of physical chance, and expert judgement where appropriate. I will expand on this below but the basic idea is that finding limitations in the evidence base will alter the width and skew of the certainty range. The width changes when we consider a wider interval to be more probable, and skew changes if we have information about the direction with which the interval is likely to change. Figure 9.3 visually demonstrates this process.

Certainty ranges are potentially a better approach to including effect sizes than using

$P(C_\varepsilon)$, because of a worry about decision making on evidence bases with quality issues. One challenge for the critical threshold approach is being able to say something about how changes in one's confidence in an effect estimate, affect decisions made using the critical threshold approach. Plausibly, a decrease in confidence in the correctness of an estimate that surpasses a critical threshold calls into doubt whether the true effect is also above that threshold. Consider again the example from above, where an interval estimate I for drug D is [1.5,2.5], and the δ for D is 0.25. In this case it seemed like this interval estimate is sufficient to recommend D. But suppose further that one uses the GRADE framework to rate the quality of evidence and discovers a serious limitation on one of the moderating domains. On the current proposal, one would now hold low confidence that I is correct, in the sense that the true value for the effectiveness of D is in I. One might be tempted to say that the true effect is lower than δ. But without some measure of the decrease in one's confidence, one cannot say this. It might be the case that the true value for the effect is not in I but is greater than δ. So, it would be better to have some idea of the likelihood that other values could be the true effect. This is what the certainty range approach does. By widening and/or skewing the certainty range, values outside of the original interval estimate are made more probable relative to values inside the estimate. In effect, the quantitation of uncertainty associated with each moderating domain leads to an adjustment of the lower bound of the interval estimate (for a visual representation of how this works see fig. 9.4). On the other hand, an approach that merely assigns a probability to C_ε, while simpler, provides no indication of what alternative values for the effect size are probable. As it allows better decisions to be made, the certainty range approach is arguably a better way to including effect size information.

Worries about certainty ranges

The certainty range approach promises a way to include effect sizes and keep the critical threshold approach, but it is not without its own problems. One charge is arbitrariness.

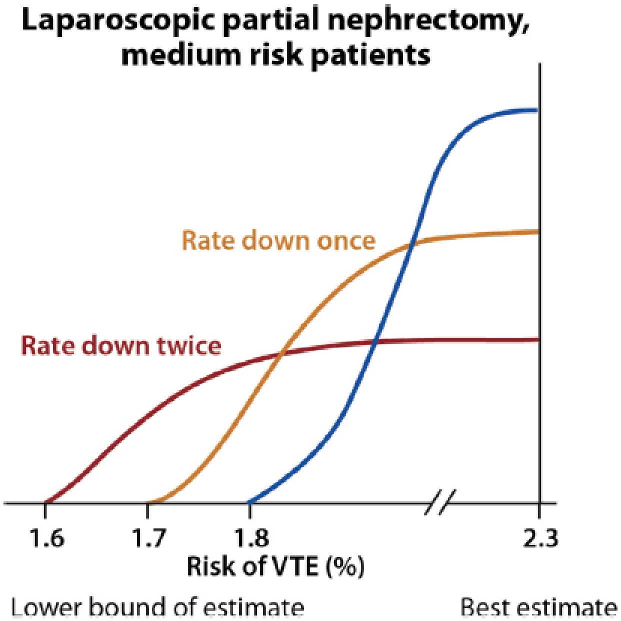


Figure 9.4: Visual illustration of the adjustment of certainty ranges from Tikkinen et al. (2017). The blue, yellow and ref lines all represent one half of a probability distribution defined over the intervval estimate. If the evidence has no quality issues then it spans the entirety of the estimate (blue line). On finding quality issues (yellow and red line) the range is adjusted, making lower values for the estimate more probable relative to higher values.

This charge was made in Chapters 6 and 7 against the use of subjective probabilities and formalisation of evidence evaluation, respectively. Tikkinen et al. (2018) also recognise that the values they assign to the certainty ranges are somewhat arbitrary. In response to worries about arbitrariness I have argued that there are ways to avoid it. One way is to make one's assumptions and reasons for assignments of values explicit. Another is to use constraints on belief that seek to limit subjectivity. These same methods could be applied to the certainty range approach. However, the use of the certainty range in Tikkinen et al. (2018) was more in the vein of a concept paper, where they intended to illustrate how it could be used. As this approach is in the early stages of development, it would be too much to ask for a complete reduction of arbitrariness. Moreover, without more detail on the mechanics of adjusting certainty ranges it is hard to say anything concrete about how one would apply my solutions to reduce arbitrariness. It suffices for the purposes of this chapter to say that the applicability of formal models of belief to evidence evaluation, which make use of probabilistic analysis tools and techniques, shows how it is possible to constrain subjectivity and arbitrariness within a formal model. It is plausible that such positives will carry over to the certainty range approach.

Another worry about arbitrariness has to do with the precision of new lower bounds for the interval estimate. I have acknowledged that complete precision in the application of formal models may be hard to achieve. This may mean that when an altered interval estimate is close to but above the δ for that intervention, then there is a distinct possibility that the interval may indeed cross that δ. One could opt for a stipulation that when a lower bound and δ are a certain distance from one another then one cannot make a decision. But again, the magnitude of this distance is likely to be arbitrary. Ultimately, this will have to come down to expert judgement. There is likely not going to be some completely agent independent way to make judgements on the strength of evidence or how to make decisions on the basis of that evidence. As suggested in §7.3.2, we should attempt to approximate formal ideals, and explicitly evaluate the extent of subjective choice required to make judgements.

A positive case for this approach can also be made by appealing to the importance of alteration of effect estimates elsewhere in medicine. *Correction factors* are used in physics to multiply the results of studies to adjust the effect of known systematic biases (e.g., Li et al. (1996); Gruttmann and Wagner (2001)). They are also used in the biomedical sciences to adjust the results of laboratory studies (Burnett and Noonan, 1974). In clinical science, they can be used to correct results for particular biases. For example, Spix et al. (2016) outline correction factors for self-selection bias in screening programs. More generally, Stegenga (2018, Appendix 5) introduces a way to correct for the effect of many kinds of bias on effect estimates obtained in clinical studies. He does so in the context of making his arguments for the medical nihilism thesis, which I have rejected. In my response to the medical nihilism thesis, I did not however deny that bias was a problem. Instead we should always try to acknowledge its potential and where it actually occurs. Stegenaga proposes to take bias into account by using correction factors to subtract from the estimate of effect. The factors correspond to quantitative estimates of the effect of bias, which can be derived from studies on the effect of particular biases (e.g., Saltaji et al. (2018); see §5.4.2 for details). To be specific, Stegenga identifies six biases which are used as correction factors: confirmation bias (C), instrument bias (I), recruitment bias (R), analysis bias (A), publication bias (P), and 'all other biases' (O). For observed effect size $\bar{x}$, and the true effect size θ:

$$\theta = \bar{x} - C - I - R - A - P - O$$

Note that Stegenga assumes that, in the main, biases in medicine will over-estimate effects. The bias correction factors thus subtract from the observed effect. This is in keeping with his medical nihilism, but it is still possible that biases may underestimate effects. So a fuller development might include equivalent terms that add to the observed effect where relevant. Overall, the presence of approaches to correcting estimates provides motivation for the certainty range approach. Concerns about the impact of bias on medical research leads to a need for techniques that allows one to adjust effect estimates. Importantly, instead of rejecting results from studies with concerns of bias, we can still use this evidence to inform decisions. This is why implementing the certainty range approach in a combined evaluation of GRADE and EEEM is an important avenue to pursue.

The presence of other ways of adjusting effect sizes might lead one to doubt that developing and utilising the GRADE proposal for a combined evaluation with EEMM is wrong headed. For one, a probabilistic proposal may be more difficult to implement than a correction factor approach. However, a bias correction approach would miss out on lots of valuable information about the strength of evidence. Recall that there is more to the strength of evidence than just risk of bias. Stegenga might say that all of the other ways that the informativeness of an evidence base is captured may fall into his O correction factor. However, often we won't have precise quantitative information for what effects this kind of feature may have on an effect size. So an adjustment that uses confidence, possibly using a formal model in the Bayesian tradition, would have more tools at its disposal to deal with this issue.

Another issue is compatibility with an evaluation that includes stability of confidence. Thus far, the certainty range is developed with the idea in mind that one holds a degree of confidence in whether the parameter is close to a point estimate, which in turn is in some sense in the interval estimate. But I have argued that one must also evaluate one's stability of confidence. One might be tempted to just take the 'widening' of the certainty range as indication of changes in stability. But this would be more akin to the imprecise probabilist's notion of spreading probabilities over a wider range. Whereas I have argued we should be approximating a precise probability model. There is however something to take from this first proposal. The widening of the probability distribution does indicate instability. But it is the new, flattened, distribution that marks a lower bound of the extent of stability. This would result in a number of different distributions. One distribution would represent the 'actual' degree of confidence one holds. Then there would be others that represent the limits of possible degrees of confidence. Such limits indicate the extent of the stability of one's confidence. For example, suppose that initially one holds 95% confidence in the correctness of an estimate $I = [a, b]$. One's certainy range would thus be a probability distribution where P$=0.95$ is spread over I. Suppose further that a quality issue is detected. The probability distribution representing the certainty range would change, and the quality issue might indicate that the lower range of one's confidence might be 50% on a future evidence base. A wider, flatter, certainty range would then represent the distribution of 50% probability over I. Suppose further that this range distributes 95% probability over $I' = [a-2, b+2]$, making values outside I more probable than initially was the case. But this would represent the stability of one's confidence, while the original certainty range represents one's degree of confidence. One could use I' to inform decisions as one might be reluctant to recommend a treatment if I' contained a value equal to δ.

One might think that it is a problem that any one evaluation is working with a number of different certainty ranges. On the current proposal one has an original distribution, plus at least two distributions that represent the limits of changes in the stability of confidence. One might worry that this makes the evaluation needlessly complex. However, this kind of iterative process is already found in the original certainty range approach. One is considering the original plus altered certainty ranges. In fact, a decision making process that makes use of all of the certainty ranges will be more informed than one that just includes an altered range representing one's degree of confidence. An example of the way multiple certainty ranges can span an interval all at once can be found in both figs. 9.2 and 9.3, and shows what the representation of both degree and stability of confidence would look like.

Finally, one might wonder how certainty ranges work for mechanism claims, both for the claim itself and for their combination with correlation claim certainty ranges to produce a certainty range for the causal claim. The causal claim should include effect size information, and it would be beneficial to use the certainty range approach: after all, decisions will be made on the causal, not correlation claim. The mechanism claim must account for effect size, but it is hard to say that a probability distribution is ever defined over an estimate of effect for each mechanism claim. Instead we just hold confidence in a proposition: that the mechanism exists (and can account for the correlation and its extent). Moreover, mechanistic studies do not typically provide information about effect sizes. However, in EEMM, there is a suggestion on how to account for the magnitude of the correlation when evaluating the mechanism claim. When evaluating the status of a mechanism claim, one must evaluate it separately on the basis of clinical studies and mechanistic studies, and then combine these into a status for the causal claim. The evidence for the mechanism obtained by means of clinical studies may include effect size information. Indeed, Parkkinen et al. (2018b, p.91) asks evaluators to consider whether "clinical study evidence [is] strong enough to make it plausible that there is a mechanism that can account for the size of the correlation". So it is plausible that when evaluating the evidence from clinical studies for a mechanism claim, one must consider the effect size information. This suggests that one should define the certainty range for the mechanism claim over the effect size derived from the clinical studies. The mechanism existence claim will be implicit: when working out the certainty range, one is working out confidence in both the existence of the mechanism and whether it can account for the specific effect size. Following the process for evaluating the status of causal claims in EEMM, the certainty range for the causal claim is then arrived at by simply taking the minimum of the certainty ranges of the correlation and mechanism claims. For example, if evidence from clinical studies suggests that the effect size is equal to $I = [0.4,0.8]$, the correlation claim leads to holding $b = 0.75$ in I, and the mechanism claim leads to holding $b = 0.5$ in I, then for the causal claim one would hold $b = 0.5$ in I. This process would have the virtue that the certainty ranges for correlation, mechanism and causal claim are all defined over the same thing: an existence claim (of correlation, mechanism or causation) and an associated magnitude for the effect of the putative casual relation.

9.4 Conclusion

I have argued that GRADE can be seen as a clinical trial evaluator, and can be used to help establish correlation claims. In doing so I defended GRADE against some critiques. To be able to integrate GRADE into the EEMM process, some alterations and extensions to the process are required. One, is that to properly arrive at a status of a correlation claim, more must be said about what it takes to establish a correlation. It is not enough to leave it solely up to a framework such as GRADE, nor is it enough to make it a purely statistical question. If EBM+ is to properly evaluate causal claims by combining evaluations of evidence of correlation and mechanism, then it must be extended to include the evaluation of correlation claims. The other alterations and extensions have to do with effect size information. Estimates of effect are central to GRADE, and I argued that they should also be more explicitly included in an EEMM evaluation. This is mainly because of their utility in a decision making process that makes use of *critical thresholds*. I then considered an extension of EEMM that utilises a proposed extension to the GRADE process, namely *certainty ranges*. Such an approach promises to make decision making more precise by quantifying the affect of quality assessments on estimates of effect. I addressed some potential problems and motivated some ways to implement this approach. As is the case for GRADE, applying the certainty range approach to EEMM is in need of operationalisation. There are likely many more questions to answer, and I consider some in §10.4.

Chapter 10

Conclusions, consequences, and future directions

10.1 Introduction

In this concluding chapter, I sum up the defence I have made for my thesis. In §10.2, I summarise the contribution of each chapter to the defence of the thesis, and the specific alterations I have proposed that I think could improve the EEMM and GRADE frameworks. I then discuss some themes that have run through the thesis, before discussing some consequences of both this thesis and the wider EBM+ approach to evaluation of causality in medicine (§10.3). I finish by considering some open questions relevant to this thesis (§10.4).

10.2 Conclusions

This thesis has defended EBM+ on the grounds that it is a more complete epistemology of causality in medicine than EBM. EBM+ improves on EBM on the question of what counts as evidence for causality in medicine. To support this idea, Chapter 2 argued that an epistemology of causality in medicine that is evidentially pluralist is better than one that is not. The rest of the thesis defended the idea that EBM+ is at least as good as EBM on the question of how to evaluate evidence. To be specific, I carried out an extended analysis of the EEMM framework, which is built on EBM+ principles. That analysis demonstrated a basic kind of feasibility (Chapter 3), defended EEMM against practical challenges (Chapter 4), defended the core concepts and processes of EEMM (Chapter 5), and countered worries about subjectivity (Chapters 6 to 8). The EEMM framework is not however a finished article, and I indicated ways to improve it (Chapters 8 and 9). These final chapters show how EBM+ can allow the kind of iterative improvement that has been a central feature of EBM (e.g., major organisations progressing from using QATs to GRADE for assessing evidence (§5.5)), and demonstrates how philosophical and conceptual work can continue to improve medical methods.

Alongside the specific chapters in Part IV that proposed ways to extend and improve EEMM, I also proposed alterations to both EEMM and GRADE in other chapters. I sum them up here. In §3.4.1 I presented the raw details of my evaluation of mechanistic studies in summary tables 3.3 to 3.5. While EEMM provides tools to extract data, these tools do not contain fields that match the guidance given for evaluating the quality of mechanistic studies found in Chapter 6 of EEMM. I constructed my own summary tables to match the guidance, which were useful in providing an intermediary step between evaluation of the studies and presentation of the narrative review. Similar summary tables would be a useful addition to the process. In Chapter 4, I proposed three specific alterations. Firstly, EEMM should provide guidance on what to do when a correlation is not established.

I proposed that when the status of a correlation is below *speculative* and above *ruled out*, one should still continue to evaluate evidence of mechanism. If the status of the mechanism claim is then found to be above *speculative*, the outcome of an evaluation should be a recommendation to pursue further research that can provide evidence for the correlation claim. The second alteration is a recommendation for the wider field to provide standardised methodological guidance for methods used in the basic sciences. This could be something for the EBM+ consortium to pursue, but would be beneficial for all working in this field. The third alteration from Chapter 4 is that EEMM should include information on reasoning strategies to use when making judgements on complexity and incompleteness. I indicated that a good place to start would be the work on discovery of mechanisms (e.g., Craver and Darden (2013)). Chapter 7 argued that formal models can help to mitigate concerns about the influence of subjective choice on an evidence evaluation by the EEMM framework. I found in favour of a precise objective Bayesian model, so EEMM should look to expand and develop their extant probabilistic interpretation. In §7.3.2, I argued that a concurrent evaluation of subjective choice should be carried out for each evidence evaluation. This would be a helpful inclusion in EEMM, or in any evidence evaluation framework.

My alterations were not confined to EEMM. I also identified ways in which the GRADE framework can be improved on. In Chapter 5, I argued that GRADE should interpret quality of evidence in terms of stability of confidence. This would revert the framework back to its pre-2011 interpretation of quality of evidence, so would not be hard to implement. While GRADE is currently designed as a framework for evaluating evidence of causation, the evidential pluralism thesis, plus my arguments from §9.2.1, motivate the need to move to seeing GRADE as a framework for evaluating 'evidence of causation on the basis of clinical studies'. Further, in cases where evidence is not high-quality, GRADE will be evaluating 'evidence of correlation on the basis of clinical studies'. Finally, a change should be made to the part of the GRADE framework that starts evidence from randomised studies at a higher quality level than non-randomised studies (§§5.5 and 9.2.1). Instead, I suggest GRADE should calibrate degrees of confidence in effect estimates to the evidence of physical chance each study under evaluation provides. There will thus be no automatic difference in the rating of each kind of study. Assessment of quality should then proceed in the usual way, but in accord with the other changes to GRADE I have proposed above.

10.3 Consequences

While this thesis has defended evidence evaluation on EBM+, it has also dealt with some more general issues for evidence evaluation. A particularly important running theme has been the role of subjectivity in evidence evaluation. I identified subjectivity worries in §§4.3, 6.4 and 9.2.1, and noted that in each case, because of the role expert judgement plays in evidence evaluation, subjective influence is unavoidable. In most cases it is however enough to recognise where subjective influence enters into an evaluation process, and we should be implementing ways to identify and mitigate the influence of subjectivity rather than ignoring it (Chapters 7 and 8). Another running theme is that a broadly evidentially diverse approach to causal evaluation is better than a more restrictive one. It is not just important that evidence from mechanistic studies should be included alongside evidence from clinical studies, but also that evidence from all kinds of clinical studies should be considered when looking for evidence of correlation and/or causation. The kind of method used matters less than whether one can find explanations for an observed relationship other than that it is causal. This means the current reliance in medicine on evidence from high-quality RCTs is misguided, not only because we also need evidence from mechanistic studies, but also because non-randomised studies are also useful. There will be cases where such studies can establish causation, as long as they can establish both a correlation and a mechanism. Such cases may be rare, but even if such evidence may struggle to fully establish causation it may be enough to establish a correlation. The EBM+ approach thus not only supports utilising evidence of mechanism, but also evidence of correlation from a variety of sources. If one can establish a correlation on non-randomised studies, then if one can also establish a mechanism on the basis of mechanistic studies, one has established causation. So the EBM+ approach not only makes evidence of mechanism important, but also opens up a greater variety of approaches to securing evidence of causation.

Opening up a more diverse range of methods for obtaining evidence of causation has significant real-world consequences. Medicine currently faces a number of challenges where the traditional reliance on evidence obtained in multiple large-scale RCTs may be hindering progress. As identified in my MERS case study (§3.3), a major threat to public health are outbreaks of severe infectious disease. Indeed, another novel coronavirus is the cause of the currently ongoing COVID-19 pandemic. This virus has been called SARS-CoV2 virus, so it is no surprise that it is closely related to SARS-CoV, as well as being from the same genus of viruses as MERS-CoV (Xu et al., 2020). As of 03/04/2020 there are 972,640 confirmed cases with 50,325 deaths across 207 countries (figures from who.int/emergencies/diseases/novel-coronavirus-2019, which is updated daily). This presents one of the largest public health crises in peace-time history.

With respect to the evaluation of causation, there are many similarities with the MERS case. Already, evidence from mechanistic studies is being used to publicly support calls for the use of certain pharmacologic interventions that have shown promise in *in vitro* studies. In response, there have been calls for large scale RCTs to be carried out to determine the efficacy of these interventions. While the scale of the outbreak is so large that sufficiently powered trials may be able to be performed (which was not the case with MERS), they will still take some time to carry out. So this strategy may not be useful in this rapidly developing pandemic. One example of such an intervention is the anti-malarial drug *chloroquinine* (Mitjà and Clotet, 2020). Inhibition of SARS-COV2 replication has been demonstrated for chloroquinine *in vitro* (Wang et al., 2020), which has motivated use on compassionate grounds in China and Italy (Gao et al., 2020). Without ways to assess this evidence from mechanistic studies we cannot say much about conclusions drawn from it, but relying solely on RCTs delays our response. An EBM+ approach in this area may thus be useful.

Another area where it may be useful is in pharmcological risk surveillance, an area in which evidential diversity is already a focus of research (e.g., Abdin et al. (2019); De Pretis et al. (2019)). It is well known that hypertension is correlated with increased incidence of death in SARS, MERS, and COVID-19 (Stockman et al., 2006; Arabi et al., 2017; Fang et al., 2020). One might be tempted to mandate isolation for the subset of the population with hypertension. However, this would mean substantial long-term isolation that would have consequently large economic effects, as in the UK alone, around 1/3 of all adults have hypertension (Joffres et al., 2013). What we need to know is whether hypertension itself is a contributory cause to death in COVID-19 patients. Evidence from mechanistic studies seems to suggest that the explanation for this correlation is actually that one kind of hypertension drug, namely angiotensin converting enzyme (ACE) inhibitors are a putative cause (Fang et al., 2020; Zheng et al., 2020). This is because SARS-CoV2 uses angiotensin converting enzyme 2 (ACE2) receptors to bind to cells, and ACE-inhibitors up-regulate this receptor in the lungs (Fang et al., 2020). There are thus more receptors available for the virus to bind to and gain entry to lung cells. As a result, the correlation between hypertension *per se* and death might be spurious. If it turns out that it is ACE inhibitors that explain the outcome, then an easy fix is to change drugs, or isolate people where this is not possible. Overall, there would be less wide an impact than if we just isolated people with hypertension. But to get to this point we need to take seriously this evidence from mechanistic studies, as well as evaluating and integrating it with evidence of correlation. Hopefully, the conclusions of this thesis will add to the growing motivation of this approach.

10.4 Future directions

I noted above that EEMM is not the finished article. To this end I proposed alterations. There are still some open questions to consider that could not be addressed within the scope of this thesis.

One arises from consideration of the COVID-19 pandemic. This is that we know that SARS-CoV2 is closely related to SARS-CoV, so we might be able to use evidence from the SARS outbreak to inform decisions in the current crisis (Xu et al., 2020). Indeed, the suggestion that upregulation of ACE2 receptors contributes to worse clinical outcomes was suggested by evidence on SARS-CoV (Gurwitz, 2020). However, the use of this kind of evidence is not always considered. For instance, the modelling that informed early UK policy used models of influenza infection as inputs, and made no reference to evidence from the SARS outbreak (Ferguson et al., 2020). While this thesis has mainly focused on efficacy, EEMM also provides guidance on establishing effectiveness, where claims in a test population are extrapolated to make claims about a target population (Parkkinen et al., 2018b, p.5). Claims that can be successfully extrapolated are deemed *externally valid*. Mechanisms are useful for extrapolating claims and EEMM provides some guidance on how to assess evidence for effectiveness claims. Another kind of extrapolation can be made from claims about one kind of object to another, e.g., from one kind of virus to another. This kind of strategy is obviously useful in the COVID-19 pandemic: evidence on SARS could be used as evidence for COVID-19. Further investigations into the guidance on extrapolation in EEMM would thus be useful. Hopefully my analysis can provide a template for such investigations.

Other questions remain around decision making on EEMM and the integration of evaluations of different kinds of evidence. With respect to decision making, I identified in §7.3.1 that if we are to include representations of stability of confidence in evaluations, then they must also be factored into decision making. I proposed that it is the extent of stability that matters. However, even this proposal is only in the nascent stages of development. More can and must be said on this matter. With respect to evidence aggregation, I considered how GRADE and EEMM are to be properly integrated. One set of questions left open concerned the development of status criteria for correlation claims. Another set of questions left open concerned how to fully formulate certainty ranges. Specific questions here include: How to define a certainty range over a mechanism claim; how to adjust certainty ranges; how to integrate certainty ranges. GRADE are also developing this approach, so answers to both these sets of questions would likely need to involve collaboration between experts in both clinical and mechanistic studies. Overall, the EBM+ approach to evaluating causality in medicine is in its infancy relative to standard EBM. The continual development and improvement of this approach is no surprise, and I look forward to following its progression.

References

Abdin, A., Auker-Howlett, D., Landes, J., Mulla, G., Jacob, C., and Osimani, B. (2019). Reviewing the mechanistic evidence assessors E-Synthesis and EBM+: A case study of amoxicillin and Drug Reaction with Eosinophilia and Systemic Symptoms (DRESS). *Current Pharmaceutical Design*, 25(16):1866–1880.

Achinstein, P. (2001). *The Book of Evidence.* Oxford University Press, New York.

Al-Tawfiq, J. A., Momattin, H., Dib, J., and Memish, Z. A. (2014). Ribavirin and interferon therapy in patients infected with the Middle East respiratory syndrome coronavirus: an observational study. *International Journal of Infectious Diseases*, 20:42–46.

Andersen, L. and Kjaer, J. (2019). Book Review: Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures. *Journal of Evaluation in Clinical Practice*, 25(6):1226–1227.

Antilla, S., Persson, J., Vareman, N., and Sahlin, N. E. (2016). Conclusiveness resolves the conflict between quality of evidence and imprecision in GRADE. *Journal of Clinical Epidemiology*, 75:1–5.

Arabi, Y. M., Balkhy, H., Hayden, F. G., Bouchama, A., Luke, T., Baillie, J. K., Omari, A. A., Hajeer, A. H., Ph, D., Senga, M., Denison, M. R., Tam, J. S. N. V., Kerkhove, M. D. V., Fowler, R. A., and Epi, M. S. (2017). Middle East Respiratory Syndrome. *The New England Journal of Medicine*, 376(6):584–594.

Assi, S. A., Imperato, M. R., Coleman, D. J. L., Pickin, A., Potluri, S., Ptasinska, A., Chin, P. S., Blair, H., Cauchy, P., James, S. R., Zacarias-cabeza, J., Gilding, L. N., Beggs, A., Clokie, S., Loke, J. C., Jenkin, P., Uddin, A., Delwel, R., Richards, S. J., Raghavan, M., Griffiths, M. J., Heidenreich, O., Cockerill, P. N., and Bonifer, C. (2019). Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nature Genetics*, 51(January):151–162.

Auker-Howlett, D. and Wilde, M. (2019). Causality in medicine, and its relation to action, mechanisms, and probability. *Metascience*, 28(3):387–391.

Auker-Howlett, D. and Wilde, M. (2020). Reinforced reasoning in medicine. *Journal of Evaluation in Clinical Practice*, 26(2):458–464.

Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., Vist, G. E., Falck-Ytter, Y., Meerpohl, J., Norris, S., and Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, 64(4):401–406.

Barnard, D. L., Day, C. W., Bailey, K., Heiner, M., Montgomery, R., Lauridsen, L., Chan, P. K., and Sidwell, R. W. (2006). Evaluation of immunomodulators, interferons and known in vitro SARS-CoV inhibitors for inhibition of SARS-CoV replication in BALB/c mice. *Antiviral Chemistry and Chemotherapy*, 17(5):275–284.

Baseler, L., Wit, E. D., and Feldmann, H. (2016). A Comparative Review of Animal Models of Middle East Respiratory Syndrome Coronavirus Infection. *Veterinary Pathology*, 53(3):521–531.

Bechtel, W. and Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36:421–441.

Bird, A. (2018). Inference to the best explanation, bayesianism, and knowledge. In Mccain, K. and Poston, T., editors, *Best Explanations: New Essays on Inference to the Best Explanation*, pages 97–120. Oxford University Press.

Braciale, T., Hahn, Y., and B, D. (2013). The Adaptive Immune Response to Viruses. In Knipe, D. and Howley, P., editors, *Fields Virology*, chapter 9, pages 215–254. Lippincott Williams and Wilkins.

Bradley, S. (2016). Imprecise Probabilities.

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*. Oxford University Press.

Broadbent, A. (2019). *Philosophy of Medicine*. Oxford University Press, New York.

Burnett, R. W. and Noonan, D. C. (1974). Calculations and correction factors used in determination of blood pH and blood gases. *Clinical Chemistry*, 20(12):1499–1506.

Campaner, R. (2011). Understanding mechanisms in the health sciences. *Theoretical Medicine and Bioethics*, 32(1):5–17.

Canali, S. (2019). Evaluating evidential pluralism in epidemiology: mechanistic evidence in exposome research. *History and Philosophy of the Life Sciences*, 41(1):1–17.

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties*, 2(1):11–20.

Catterall, W. A. (2010). Ion channel voltage sensors: Structure, function, and pathophysiology. *Neuron*, 67(6):915–928.

Chan, J. F. W., Chan, K.-h., Kao, R. Y. T., To, K. K. W., Zheng, B.-j., Li, C. P. Y., Li, P. T. W., Dai, J., Mok, F. K. Y., Chen, H., Hayden, F. G., and Yuen, K.-y. (2013). Broad-spectrum antivirals for the emerging Middle East respiratory syndrome coronavirus. *Journal of Infection*, 67(6):606–616.

Chan, J. F. W., Yao, Y., Yeung, M. L., Deng, W., Bao, L., Jia, L., Li, F., Xiao, C., Gao, H., Yu, P., Cai, J. P., Chu, H., Zhou, J., Chen, H., Qin, C., and Yuen, K. Y. (2015). Treatment with lopinavir/ritonavir or interferon-β1b improves outcome of MERSCoV infection in a nonhuman primate model of common marmoset. *Journal of Infectious Diseases*, 212(12):1904–1913.

Channappanavar, R., Fehr, A. R., Vijay, R., Mack, M., Zhao, J., Meyerholz, D. K., and Perlman, S. (2016). Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice. *Cell Host and Microbe*, 19(2):181–193.

Chen, X., Wang, Q., Ni, F., and Ma, J. (2010). Structure of the full-length Shaker potassium channel Kv1.2 by normal-mode-based X-ray crystallographic refinement. *Proceedings of the National Academy of Sciences of the United States of America*, 107(25):11352–11357.

Cheng, K. W., Cheng, S. C., Chen, W. Y., Lin, M. H., Chuang, S. J., Cheng, I. H., Sun, C. Y., and Chou, C. Y. (2015). Thiopurine analogs and mycophenolic acid synergistically inhibit the papain-like protease of Middle East respiratory syndrome coronavirus. *Antiviral Research*, 115:9–16.

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., Leucht, S., Ruhe, H. G., Turner, E. H., Higgins, J. P. T., Egger, M., Takeshima, N., Hayasaka, Y., Imai, H., Shinohara, K., Tajika, A., Ioannidis, J. P. A., and Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet*, 391(10128):P1357–1366.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventive Medicine*, 57(6):745–747.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2014). Mechanisms and the Evidence Hierarchy. *Topoi*, 33(2):339–360.

Coen, D. and Richman, D. (2013). Antiviral Agents. In Knipe, D. and Howley, P., editors, *Fields Virology*, chapter 13, pages 338–374. Lippincott Williams and Wilkins, 6th edition.

Condit, R. C. (2013). Principles of Virology. In Knipe, D. and Howley, P., editors, *Fields Virology*, chapter 2, pages 21–51. Lippincott Williams and Wilkins, 6th edition.

Corman, V. M., Albarrak, A. M., Omrani, A. S., Albarrak, M. M., Farah, M. E., Almasri, M., Muth, D., Sieberg, A., Meyer, B., Assiri, A. M., Binger, T., Steinhagen, K., Lattwein, E., Al-Tawfiq, J., Müller, M. A., Drosten, C., and Memish, Z. A. (2015). Viral Shedding and Antibody Response in 37 Patients with Middle East Respiratory Syndrome Coronavirus Infection. *Clinical Infectious Diseases*, 62(4):477–483.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153(3):355–376.

Craver, C. F. and Darden, L. (2013). *In Search of Mechanisms: Discoveries Across the Life Sciences*. The University of Chicago Press, Chicago.

Darling, A. J., Boose, J. A., and Spaltro, J. (1998). Virus assay methods: Accuracy and validation. *Biologicals*, 26(2):105–110.

de Groot, R. J. D., Baker, S. C., Baric, R. S., Brown, C. S., Drosten, C., Enjuanes, L., and Fouchier, R. A. M. (2013). Middle East Respiratory Syndrome Coronavirus (MERS-CoV): Announcement of the Coronavirus Study Group. *Journal of Virology*, 87(14):7790–7792.

De Pretis, F., Landes, J., and Osimani, B. (2019). E-synthesis: A Bayesian framework for causal assessment in pharmacosurveillance. *Frontiers in Pharmacology*, 10(December):1–20.

de Wit, E., Rasmussen, A. L., Falzarano, D., Bushmaker, T., Feldmann, F., Brining, D. L., Fischer, E. R., Martellaro, C., Okumura, A., Chang, J., Scott, D., Benecke, A. G., Katze, M. G., Feldmann, H., and Munster, V. J. (2013). Middle East respiratory syndrome coronavirus (MERS-CoV) causes transient lower respiratory tract infection in rhesus macaques. *Proceedings of the National Academy of Sciences*, 110(41):16598–16603.

Devanesan, A. (2019). Medical nihilism: The limits of a decontextualised critique of medicine. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 79(February):101189.

Devaraj, S. G., Wang, N., Chen, Z., Chen, Z., Tseng, M., Barretto, N., Lin, R., Peters, C. J., Tseng, C. T. K., Baker, S. C., and Li, K. (2007). Regulation of IRF-3-dependent innate immunity by the papain-like protease domain of the severe acute respiratory syndrome coronavirus. *Journal of Biological Chemistry*, 282(44):32208–32221.

Dupré, J. (2013). I—John Dupré: Living Causes. *Aristotelian Society Supplementary Volume*, 87(1):19–37.

European Medical Agency (2018). Assessment report - Zirabev (EMEA/H/C/004697/0000). *Committee for Medicinal Products for Human Use*, 13(December):1–86.

Falzarano, D., Wit, E. D., Martellaro, C., Callison, J., Munster, V. J., and Feldmann, H. (2013a). Inhibition of novel b coronavirus replication by a combination of interferon-a2b and ribavirin. *Scientific Reports*, 3:1–6.

Falzarano, D., Wit, E. D., Rasmussen, A. L., Feldmann, F., Okumura, A., Scott, D. P., Brining, D., Bushmaker, T., Martellaro, C., Baseler, L., Benecke, A. G., Katze, M. G., Munster, V. J., and Feldmann, H. (2013b). Treatment with interferon-a2b and ribavirin improves outcome in MERS-CoV–infected rhesus macaques. *Nature Medicine Letters*, 19(10):1313–1317.

Fang, L., Karakiulakis, G., and Roth, M. (2020). Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? *The Lancet Respiratory*, 2600(20):30116.

Feinstein, A. R. and Horwitz, R. I. (1997). Problems in the "Evidence" of "Evidence-based Medicine". *The American Journal of Medicine*, 103:529–535.

Feld, J. J. and Hoofnagle, J. H. (2005). Mechanism of action of interferon and ribavirin in treatment of hepatitis C. *Nature*, 436(7053):967–972.

Ferguson, N. M., Laydon, D., Nedjati-gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-dannenburg, G., Dighe, A., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., Van, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P. G. T., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S., and Ghani, A. C. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID- 19 mortality and healthcare demand.

Fletcher, R., Fletcher, S., and Fletcher, G. (2012). *Clinical Epidemiology: The Essentials.* Lippincott Williams and Wilkins.

Frieman, M., Yount, B., Heise, M., Kopecky-Bromberg, S. A., Palese, P., and Baric, R. S. (2007). Severe Acute Respiratory Syndrome Coronavirus ORF6 Antagonizes STAT1 Function by Sequestering Nuclear Import Factors on the Rough Endoplasmic Reticulum/Golgi Membrane. *Journal of Virology*, 81(18):9812–9824.

Gao, J., Tian, Z., and Yang, X. (2020). Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *Bioscience trends*, 14(1):72–73.

Gillies, D. (2000). *Philosophical Theories of Probability.* Taylor and Francis, London.

Gillies, D. (2005). Hempelian and Kuhnian approaches in the philosophy of medicine: The Semmelweis case. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(1):159–181.

Gillies, D. (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*, chapter 6. Oxford University Press.

Gillies, D. (2019a). *Causality, Probability, and Medicine.* Routledge, Abingdon.

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

Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, 69(S3):S342–S353.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350.

Grossman, J. and Mackenzie, F. J. (2005). The randomized controlled trial: gold standard, or merely standard? *Perspectives in biology and medicine*, 48(4):516–534.

Gruttmann, F. and Wagner, W. (2001). Shear correction factors in timoshenko's beam theory for arbitrary shaped cross-sections. *Computational Mechanics*, 27(3):199–207.

Gurwitz, D. (2020). Angiotensin receptor blockers as tentative SARS-CoV-2 therapeutics. *Drug Development Research*, (February):2–5.

Guyatt, G., Oxman, A. D., Sultan, S., Brozek, J., Glasziou, P., Alonso-Coello, P., Atkins, D., Kunz, R., Montori, V., Jaeschke, R., Rind, D., Dahm, P., Akl, E. A., Meerpohl, J., Vist, G., Berliner, E., Norris, S., Falck-Ytter, Y., and Schünemann, H. J. (2013). GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *Journal of Clinical Epidemiology*, 66(2):151–157.

Guyatt, G. H., Oxman, A. D., Kunz, R., Atkins, D., Brozek, J., Vist, G., Alderson, P., Glasziou, P., Falck-Ytter, Y., and Schünemann, H. J. (2011a). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology*, 64(4):395–400.

Guyatt, G. H., Oxman, A. D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., Devereaux, P. J., Montori, V. M., Freyschuss, B., Vist, G., Jaeschke, R., Williams, J. W., Murad, M. H., Sinclair, D., Falck-Ytter, Y., Meerpohl, J., Whittington, C., Thorlund, K., Andrews, J., and Schünemann, H. J. (2011b). GRADE guidelines 6. Rating the quality of evidence - Imprecision. *Journal of Clinical Epidemiology*, 64(12):1283–1293.

Guyatt, G. H., Oxman, A. D., Kunz, R., Jaeschke, R., Helfand, M., Liberati, A., Vist, G. E., and Schünemann, H. J. (2008). GRADE: Incorporating considerations of resources use into grading recommendations. *BMJ (Clinical research ed.)*, 336(7654):1170–3.

Haagmans, B. L., Kuiken, T., Martina, B. E., Fouchier, R. A. M., Rimmelzwaan, G. F., Van Amerongen, G., Van Riel, D., De Jong, T., Itamura, S., Chan, K. H., Tashiro, M., and Osterhaus, A. D. M. E. (2004). Pegylated interferon-α protects type 1 pneumocytes against SARS coronavirus infection in macaques. *Nature Medicine*, 10(3):290–293.

Hájek, A. (2008). Arguments for-or against-probabilism. *British Journal for the Philosophy of Science*, 59(4):793–819.

Hart, B. J., Dyall, J., Postnikova, E., Zhou, H., Kindrachuk, J., Johnson, R. F., Olinger, G. G., Frieman, M. B., Holbrook, M. R., Jahrling, P. B., and Hensley, L. (2014). Interferon-β and mycophenolic acid are potent inhibitors of middle east respiratory syndrome coronavirus in cell-based assays. *Journal of General Virology*, 95(PART3):571–577.

Hensley, L. E., Fritz, E. A., Jahrling, P. B., Karp, C. L., Huggins, J. W., and Geisbert, T. W. (2004). Interferon-β 1a and SARS Coronavirus Replication. *Emerging Infectious Diseases*, 10(2):317–319.

Higgins, J., Thompson, S., Deeks, J., and Altman, D. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327:557–560.

Holman, B. (2018). In defense of meta-analysis. *Synthese*, 196:3189–3211.

Howick, J. (2011a). Exposing the Vanities - and a Qualified Defense - of Mechanistic Reasoning in Health Care Decision Making. *Philosophy of Science*, 78(5):926–940.

Howick, J. (2011b). *The Philosophy of Evidence-Based Medicine*. Wiley-Blackwell.

Howson, C. (2003). *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press, Oxford.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach.* Open Court Publishing, Peru; US, 3rd edition.

Hultcrantz, M., Rind, D., Akl, E. A., Treweek, S., Mustafa, R. A., Iorio, A., Alper, B. S., Meerpohl, J. J., Murad, M. H., Ansari, M. T., Katikireddi, S. V., Östlund, P., Tranæus, S., Christensen, R., Gartlehner, G., Brozek, J., Izcovich, A., Schünemann, H., and Guyatt, G. (2017). The GRADE Working Group clarifies the construct of certainty of evidence. *Journal of Clinical Epidemiology*, 87(July):4–13.

Illari, P. M. (2011). Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis. *International Studies in the Philosophy of Science*, 25(2):139–157.

Illari, P. M. K. and Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, 2(1):119–135.

Iwasaki, A. and Medzhitov, R. (2013). Innate Responses to Viral Infections. In Knipe, D. and Howley, P., editors, *Fields Virology*, chapter 8, pages 189–213. Lippincott Williams and Wilkins, 6th edition.

Joffres, M., Falaschetti, E., Gillespie, C., Robitaille, C., Loustalot, F., Poulter, N., McAlister, F. A., Johansen, H., Baclic, O., and Campbell, N. (2013). Hypertension prevalence, awareness, treatment and control in national surveys from England, the USA and Canada, and correlation with stroke and ischaemic heart disease mortality: A cross-sectional study. *BMJ Open*, 3(8):1–9.

Joyce, J. M. (2005). How Probabilities Reflect Evidence. *Philosophical Perspectives*, 19:153–178.

Joyce, J. M. (2011a). A Defence of Imprecise Credences in Inference and Decision Making. *Philosophical Perspectives*, 24:281–323.

Joyce, J. M. (2011b). The Development of Subjective Bayesianism. *Handbook of the History of Logic*, 10:415–475.

Kelly, T. (2008). Evidence: Fundamental Concepts and the Phenomenal Conception. *Philosophy Compass*, 3(5):933–955.

Keynes, J. (1921). *A Treatise on Probability.* Macmillan, London.

Kumar, A., Miladinovic, B., Guyatt, G. H., Schünemann, H. J., and Djulbegovic, B. (2016). GRADE guidelines system is reproducible when instructions are clearly operationalized even among the guidelines panel members with limited experience with GRADE. *Journal of Clinical Epidemiology*, 75:115–118.

Kyburg, H. E. (2003). Are there degrees of belief? *Journal of Applied Logic*, 1(3-4):139–149.

Kyburg Jr, H. E. and Teng, C. M. (2001). *Uncertain Inference.* Cambridge University Press.

La Caze, A. (2009). Evidence-based medicine must be ... *Journal of Medicine and Philosophy*, 34(5):509–527.

La Caze, A. (2019). Better evaluating mechanisms in medicine. Book Review: Evaluating Evidence of Mechanisms in Medicine. *Journal of Evaluation in Clinical Practice*, 25(6):1228–1231.

Landes, J., Osimani, B., and Poellinger, R. (2018). Epistemology of causal inference in pharmacology: Towards a framework for the assessment of harms. *European Journal for Philosophy of Science*, 8(1):3–49.

Lewis, D. K. (1980). A Subjectivist's Guide to Objective Chance. In Jeffrey, R. C., editor, *Studies in Inductive Logic and Probability Vol II*, pages 263–293. University of California Press, Berkeley.

Li, G., Abrahams, A. D., and Atkinson, J. F. (1996). Correction factors in the determination of mean velocity of overland flow. *Earth Surface Processes and Landforms*, 21(6):509–515.

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

Masters, P. S. and Perlman, S. (2013). Coronaviridae. In Knipe, D. M. and Howley, P. M., editors, *Fields Virology*, chapter 28, pages 825–858. Lippincott Williams and Wilkins, 6th edition.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. The University of Chicago Press, London.

McCain, K. and Poston, T. (2014). Why Explanatoriness Is Evidentially Relevant. *Thought*, 3:145–153.

Menachery, V. D., Mitchell, H. D., Cockrell, A. S., Gralinski, L. E., Yount, B. L., Graham, R. L., McAnarney, E. T., Douglas, M. G., Scobey, T., Beall, A., Dinnon, K., Kocher, J. F., Hale, A. E., Stratton, K. G., Waters, K. M., and Baric, R. S. (2017). MERS-CoV accessory orfs play key role for infection and pathogenesis. *mBio*, 8(4):1–14.

Mercuri, M., Baigrie, B., and Upshur, R. E. (2018). Going from evidence to recommendations: Can GRADE get us there? *Journal of Evaluation in Clinical Practice*, 24(5):1232–1239.

Mercuri, M. and Gafni, A. (2018a). The evolution of GRADE (part 1): Is there a theoretical and / or empirical basis for the GRADE framework ? *Journal of Evaluation in Clinical Practice*, 24(5):1203–1210.

Mercuri, M. and Gafni, A. (2018b). The evolution of GRADE (part 2): Still searching for a theoretical and / or empirical basis for the GRADE framework. *Journal of Evaluation in Clinical Practice*, 24(5):1211–1222.

Mercuri, M. and Gafni, A. (2018c). The evolution of GRADE (part 3): A framework built on science or faith ? *Journal of Evaluation in Clinical Practice*, 24(5):1223–1231.

Milne-Price, S., Miazgowicz, K. L., and Munster, V. J. (2014). The emergence of the Middle East respiratory syndrome coronavirus. *Pathogens and Disease*, 71:121–136.

Mitjà, O. and Clotet, B. (2020). Use of antiviral drugs to reduce COVID-19 transmission. *The Lancet Global health*, (March):1–2.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review*, 23(1):103–123.

Morgenstern, B., Michaelis, M., Baerb, P. C., Doerra, H. W., and Cinatl Jr, J. (2005). Ribavirin and interferon-b synergistically inhibit SARS-associated coronavirus replication in animal and human cell lines. *Biochemical and Biophysical Research Communications*, 326:905–908.

Mustafa, R. A., Santesso, N., Brozek, J., Akl, E. A., Walter, S. D., Norman, G., Kulasegaram, M., Christensen, R., Guyatt, G. H., Falck-Ytter, Y., Chang, S., Murad, M. H., Vist, G. E., Lasserson, T., Gartlehner, G., Shukla, V., Sun, X., Whittington, C., Post, P. N., Lang, E., Thaler, K., Kunnamo, I., Alenius, H., Meerpohl, J. J., Alba, A. C., Nevis, I. F., Gentles, S., Ethier, M. C., Carrasco-Labra, A., Khatib, R., Nesrallah, G., Kroft, J., Selk, A., Brignardello-Petersen, R., and Schünemann, H. J. (2013). The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *Journal of Clinical Epidemiology*, 66(7):736–742.

National Institute for Clinical Excellence (2012). The Guidelines Manual.

Okasha, S. (2002). Darwinian metaphysics: Species and the question of essentialism. *Synthese*, 131(2):191–213.

Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., Norell, C., Russo, F., Shaw, B., and Williamson, J. (2018a). Evaluating Evidence of Mechanisms in Medicine (Appendices).

Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., Norell, C., and Williamson, J. (2018b). *Evaluating evidence of mechanisms in medicine: principles and procedures*. Springer.

Pierson, T. C. and Diamond, M. S. (2013). Flaviviruses. In Knipe, D. and Howley, P., editors, *Fields Virology*, chapter 28, pages 747–794. Lippincott Williams and Wilkins, 6 edition.

Plutynski, A. (2018). *Explaining Cancer: Finding Order in Disorder*. Oxford University Press, New York.

Pogue, J. M. and Yusuf, S. (1997). Cumulating evidence from randomized trials: Utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials*, 18(6):580–593.

Rabaan, A. A., Alahmed, S. H., Bazzi, A. M., and Alhani, H. M. (2017). A review of candidate therapies for Middle East respiratory syndrome from a molecular perspective. *Journal of Medical Microbiology*, 66:1261–1274.

Roche, W. and Sober, E. (2013). Explanatoriness is evidentially irrelevant, or inference to the best explanation meets Bayesian confirmation theory. *Analysis*, 73(4):659–668.

Roldão, A., Oliveira, R., Carrondo, M. J. T., and Alves, P. M. (2009). Error assessment in recombinant baculovirus titration : Evaluation of different methods. *Journal of Virological Methods*, 159:69–80.

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

Sackett, D., Rosenberg, W., Gray, J., Haynes, R., and Richardson, W. (1996). Evidence based medicine: what it is and what it isn't. *British medical Journal*, 312(1):71–72.

Saltaji, H., Armijo-Olivo, S., Cummings, G., Amin, M., da Costa, B., and Flores-Mir, C. (2018). Impact of Selection Bias on Treatment Effect Size Estimates in Randomized Trials of Oral Health Interventions: A meta-epidemiological Study. *Journal of Dental Research*, 97(1):5–13.

Schünemann, H. J. (2016). Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? *Journal of Clinical Epidemiology*, 75:6–15.

Skyrms, B. (1977). Resiliency, Propensities, and Causal Necessity. *The Journal of Philosophy*, 74(11):704–713.

Smith, M. T., Guyton, K. Z., Gibbons, C. F., Fritz, J. M., Portier, C. J., Rusyn, I., DeMarini, D. M., Caldwell, J. C., Kavlock, R. J., Lambert, P. F., Hecht, S. S., Bucher, J. R., Stewart, B. W., Baan, R. A., Cogliano, V. J., and Straif, K. (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. *Environmental Health Perspectives*, 124(6):713–721.

Sober, E. (2001). Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause. *The British Journal for the Philosophy of Science*, 52:331–346.

Solomon, M. (2015). *Making Medical Knowledge*. Oxford University Press, New York.

Solomon, M. (2019). Review of Anya Plutynski's Explaining cancer: finding order in disorder. *Biology and Philosophy*, 34(3):1–6.

Spix, C., Berthold, F., Hero, B., Michaelis, J., and Schilling, F. H. (2016). Correction factors for self-selection when evaluating screening programmes. *Journal of Medical Screening*, 23(1):44–49.

Sprenger, J. (2018). The Objectivity of Subjective Bayesian Inference. *European Journal for Philosophy of Science*, 8:539–588.

Staatz, C. E. and Tett, S. E. (2014). Pharmacology and toxicology of mycophenolate in organ transplant recipients: An update. *Archives of Toxicology*, 88(7):1351–1389.

Staley, K. W. (2017). Pragmatic warrant for frequentist statistical practice: the case of high energy physics. *Synthese*, 194(2):355–376.

Stegenga, J. (2013). Evidence in biology and the conditions of success. *Biology and Philosophy*, 28:981–1004.

Stegenga, J. (2018). *Medical Nihilism*. Oxford University Press, Oxford.

Stockman, L. J., Bellamy, R., and Garner, P. (2006). SARS: Systematic Review of Treatment Effects. *PLoS Medicine*, 3(9):1525–1531.

Thornton, J., Alderson, P., Tan, T., Turner, C., Latchem, S., Shaw, E., Ruiz, F., Reken, S., Mugglestone, M. A., Hill, J., Neilson, J., Westby, M., Francis, K., Whittington, C., Siddiqui, F., Sharma, T., Kelly, V., Ayiku, L., and Chamberlain, K. (2013). Introducing GRADE across the NICE clinical guideline program. *Journal of Clinical Epidemiology*, 66(2):124–131.

Tijms, H. (2012). *Understanding Probability.* Cambridge University Press, Cambridge, 3rd edition.

Tikkinen, K. A., Craigie, S., Schünemann, H. J., and Guyatt, G. H. (2018). Certainty ranges facilitated explicit and transparent judgments regarding evidence credibility. *Journal of Clinical Epidemiology*, 104:46–51.

Tikkinen, K. A. O., Cartwright, R., Gould, M. K., Naspro, R., Novara, G., Sandset, P. M., Violette, P. D., and Guyatt, G. H. (2017). European Association of Urology Guidelines on Thromboprophylaxis in Urological Surgery. *EAU Guideline Office.*

Von Mises, R. (1982). *Probability, Statistics and Truth.* Dover, 2nd edition.

Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., Shi, Z., Hu, Z., Zhong, W., and Xiao, G. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Research*, 30:269–271.

White, R. (2010). Evidential Symmetry and Mushy Credence. *Oxford Studies in Epistemology*, 3:161–186.

Wilde, A. H. D., Raj, V. S., Oudshoorn, D., Bestebroer, T. M., Nieuwkoop, S. V., Limpens, R. W. A. L., Posthuma, C. C., Meer, Y. V. D., Haagmans, B. L., Snijder, E. J., and Hoogen, B. G. V. D. (2013). MERS-coronavirus replication induces severe in vitro cytopathology and is strongly inhibited by cyclosporin A or interferon- a treatment. *Journal of General Virology*, 94:1749–1760.

Williamson, J. (2010). *In Defence of Objective Bayesianism.* Oxford University Press.

Williamson, J. (2019). Establishing Causal Claims in Medicine. *International Studies in the Philosophy of Science*, 32(1):33–61.

Williamson, J. (2020). The feasibility and malleability of EBM+. Unpublishe:1–16.

Worrall, J. (2002). What evidence in evidence based medicine? *Philosophy of Science*, 69(September):316–330.

Worrall, J. (2007). Evidence in Medicine and Evidence-Based Medicine. *Philosophy Compass*, 26(10):981–1022.

Xu, J., Zhao, S., Teng, T., Abdalla, A. E., Zhu, W., Xie, L., Wang, Y., and Guo, X. (2020). Systematic comparison of two animal-to-human transmitted human coronaviruses: SARS-CoV-2 and SARS-CoV. *Viruses*, 12(2).

Yang, Y., Ling, Z., Heyuan, G., Yao, D., Baoying, H., Yin, G., Zhengdong, Z., and Wenjie, T. (2013). The structural and accessory proteins M , ORF 4a , ORF 4b , and ORF 5 of Middle East respiratory syndrome coronavirus (MERS-CoV) are potent interferon antagonists. *Protein & Cell*, 4(12):951–961.

Yu, P., Xu, Y., Deng, W., Bao, L., Huang, L., and Xu, Y. (2017). Comparative pathology of rhesus macaque and common marmoset animal models with Middle East respiratory syndrome coronavirus. *PLoS ONE*, 12(2):e0172093.

Zheng, Y.-Y., Ma, Y.-T., Zhang, J.-Y., and Xie, X. (2020). COVID-19 and the cardiovascular system. *Nature reviews. Cardiology.*

Zumla, A., Hui, D. S., and Perlman, S. (2015). Middle East respiratory syndrome. *The Lancet*, 386(9997):995–1007.