Chapter 1
Evaluating evidence of mechanisms in medicine
Principles and procedures

Veli-Pekka Parkkinen, Christian Wallmann, Michael Wilde, Brendan Clarke, Phyllis Illari, Michael P. Kelly, Federica Russo, Beth Shaw and Jon Williamson

1.1 Introduction ................................................................. 3
1.1.1 What is a mechanism? ............................................... 4
1.1.2 Why consider evidence of mechanisms? ......................... 4
1.2 How to consider evidence of mechanisms: an overview ................. 9
1.2.1 Questions to address ................................................ 9
1.2.2 Quality of evidence and status of claim ......................... 11
1.2.3 Overall approach .................................................... 12
1.3 Gathering evidence of mechanisms .................................... 18
1.3.1 Identify specific mechanism hypotheses ......................... 19
1.3.2 Formulate the review questions ................................... 22
1.3.3 Search the literature ................................................ 24
1.3.4 Refine results of the search ...................................... 24
1.3.5 Presenting the evidence of mechanisms ......................... 26
1.4 Evaluating evidence of mechanisms .................................. 27
1.4.1 Overview ............................................................... 27
1.4.2 Evaluating mechanistic studies ................................... 29
1.4.3 Determining the status of the general mechanistic claim ........ 32
1.4.4 Presenting the quality of evidence of mechanisms ............ 36
1.5 Using evidence of mechanisms to evaluate efficacy and external validity .......... 39
1.5.1 Efficacy ............................................................... 39
1.5.2 External validity ..................................................... 42
1.5.3 Presenting the status of a causal claim ............................ 45
1.A A critical appraisal tool for evidence of mechanisms .................. 46

Veli-Pekka Parkkinen · Christian Wallmann · Michael Wilde · Jon Williamson
Centre for Reasoning, University of Kent

Brendan Clarke · Phyllis Illari
Department of Science and Technology Studies, University College London

Michael P. Kelly
Department of Public Health and Primary Care, University of Cambridge

Federica Russo
Department of Philosophy, University of Amsterdam

Beth Shaw
National Institute of Health and Care Excellence
Abstract This chapter presents heuristics for assessing evidence of mechanisms when evaluating effectiveness in medicine. The problem of evaluating effectiveness is broken down into two sub-problems: evaluating efficacy and evaluating external validity. In each case, we develop a three-step procedure for handling evidence of mechanisms: (i) collecting evidence of mechanisms; (ii) evaluating evidence of mechanisms; (iii) combining this evaluation with an evaluation of evidence of correlation in order to produce an overall assessment of causality. Appendices provide appraisal tools to assist with the assessment of evidence of mechanisms and discuss particular applications.

Key words: Evidence-based medicine, mechanisms, effectiveness, efficacy, external validity, causation, evidence appraisal

This chapter is primarily aimed at guideline developers and those charged with evidence appraisal, in the context of evaluating the effectiveness of proposed interventions on health or the effectiveness of certain environmental exposures on health.

At present, evidence of mechanisms tends to influence the evaluation of effectiveness in an implicit way, mediated by the opinions of experts, particularly expert panel members on evidence appraisal committees. As explained below, evidence of mechanisms can often provide very good evidence of effectiveness, so it is right that it should be taken into account. However, the lesson of evidence-based medicine is that one needs to make evidence explicit in order to scrutinise and challenge it properly, and that one needs to make explicit the ways in which evidence is evaluated in order to improve these methods of evaluation. This chapter seeks to extend this evidence-based approach to evidence of mechanisms.
1 Evaluating evidence of mechanisms in medicine

1.1 Introduction

Much of medical practice\(^1\) depends upon establishing causal claims—i.e., on establishing effectiveness.\(^2\) This includes, for example, identifying the causes of cancers in humans, evaluating whether a medical device will lead to improved outcomes in a particular patient, establishing whether a public health action will have the desired effects in the target population, and ascertaining the cost effectiveness of a health intervention on a target population.

Establishing effectiveness goes well beyond determining whether there is a correlation between two variables of interest in a study population. One needs to establish two further claims:

1. That this correlation is genuinely causal in the study population. This is often called efficacy,\(^3\) or the internal validity of the causal relationship. The study population is typically a population that satisfies certain ideal experimental conditions or that experiences a narrowly circumscribed range of exposures.

2. That this causal relationship applies outside the study population to the population of interest. This is often called applicability, or the external validity of the causal relationship. The target population of interest may be the population of all patients with a particular condition or a population of all individuals exposed to an environmental agent, for instance. Typically, many individuals in this target population will not satisfy the ideal conditions that characterise the study population.

As we shall discuss below, evidence of mechanisms is crucial to both of these steps. Evidence of mechanisms does not exhaust the evidence for these causal inferences, but it contributes to them in important ways (Russo & Clarke, forthcoming).

First, we shall explain what we mean by ‘mechanism’.

---

\(^1\) With the term ‘medical practice’ we shall understand, broadly, clinical, scientific, and political forms of engagement with health and disease (Clarke and Russo, 2016). This comprises clinical practice, including primary care and hospital medicine, preventive medicine and public health, and epidemiology. This inclusive perspective covers approaches as diverse as evidence-based practice, guideline development, personalised medicine, narrative medicine, and others.

\(^2\) Definitions of terms that are underlined are included in the Glossary, Appendix 1.H.

\(^3\) Sometimes, ‘efficacy’ is used only with regard to a beneficial effect of an intervention, rather than a harm such as a detrimental side-effect of a drug, or the effects of an environmental exposure. Here we use ‘efficacy’ in the more general sense to denote any sort of causal relationship within the confines of a study population. Sometimes, a distinction is drawn between ‘hazard’ and ‘risk’. Determining whether A is a cause of B is a question of hazard. Once this qualitative question is answered affirmatively, the question of risk arises, i.e., the quantitative question as to the extent of the effect size. Here we focus on the qualitative question—i.e., hazard rather than risk.
1.1.1 What is a mechanism?

Mechanisms are invoked to explain (Machamer et al., 2000). Textbooks in the biomedical and social sciences are replete with diagrams and descriptions of mechanisms. These are used to explain the proper function of features of the human body, to explain diseases and their spread, to explain the functioning of medical devices, and to explain social aspects of health interventions, among other things.

One kind of mechanism, a complex-systems mechanism, is a complex arrangement of entities and activities, organised in such a way as to be regularly or predictably responsible for the phenomenon to be explained (Illari and Williamson, 2012). In such mechanisms, spatio-temporal and hierarchical organisation tend to play a crucial explanatory role (c.f., Figure ??).

Another kind of mechanism, a mechanistic process, consists in a spatio-temporal pathway along which certain features are propagated from the starting point to the end point (Salmon, 1998). Examples include the motion of a billiard ball from cue to collision, and the trajectory of a molecule in the bloodstream from injection to metabolism. This sort of mechanism is often one-off, rather than operating in a regular and repeatable way. In the case of environmental causes of disease, the repercussions of these processes may take a long time to develop—e.g., they may be mediated by epigenetic changes.

In the health sciences, mechanistic explanations often involve a combination of these two sorts of mechanism. For example, an explanation of a certain cancer may appeal to the mechanistic processes that bring environmental factors into the human body, the eventual failure of the body’s complex-systems mechanisms for preventing damage, and the resulting mechanistic processes that lead to disease, including the propagation of tumours (Russo and Williamson, 2012).

We shall use ‘mechanism’ to refer to a complex-systems mechanism or a mechanistic process or some combination of the two. We should emphasise that mechanisms in medicine and public health may be social as well as biological (see Appendix 1.E), and, in the case of medical devices for instance, they may also include technological components.

While performing a clinical study is the usual method for establishing that two variables are correlated, a much wider variety of methods can provide good evidence of mechanisms—see Table ?? . A mechanistic study is a study which provides evidence of the details of a mechanism.

1.1.2 Why consider evidence of mechanisms?

There are various reasons for taking evidence of mechanisms into account when assessing claims in medicine. In general, when evidence is limited, the more evidence one can take into account, and the more varied this evidence is, the more reliable the resulting assessments (Landes, 2017). Moreover, when deciding whether to approve a new health intervention, or whether a chemical is carcinogenic, for example, it can...
take a very long time to gather enough evidence if the only evidence one considers is clinical study evidence. By considering evidence of mechanisms in conjunction with clinical study evidence, decisions can be made earlier: one can reduce the time taken for a drug to reach market, and reduce the time taken to restrict exposure to carcinogens, for instance.

There are also reasons for considering evidence of mechanisms that are particular to the task at hand. While evidence of mechanisms can inform a variety of tasks (see below), in this chapter we focus on its use for evaluating efficacy and external validity. Chapter 2 provides a detailed justification of the need for evidence of mechanisms when performing these two tasks. Here we shall briefly sketch the main considerations.

**Evaluating efficacy**

Establishing effectiveness can be broken down into two steps: establishing efficacy and establishing external validity. Establishing efficacy, i.e., that \( A \) is a cause of \( B \) in the study population, in turn requires establishing two things. First, \( A \) and \( B \) need to be appropriately correlated. Second, this correlation needs to be attributable to \( A \) causing \( B \), rather than some other explanation—see Table ??.

If it is genuinely the case that \( A \) is a cause of \( B \), then there is some combination of mechanisms that explains instances of \( B \) by invoking instances of \( A \) and which can account for the magnitude of the observed correlation. Thus, in order to establish efficacy one needs to establish both the existence of an appropriate correlation and the existence of an appropriate mechanism that can explain that correlation. We shall refer to this latter claim as the general mechanistic claim for efficacy.

More generally, evidence of mechanisms can help rule in or out many of the possible explanations listed in Table ???. For example, it can help to determine the direction of causation, which variables are potential confounders, whether a treatment regime is likely to lead to performance bias, and whether measured variables are likely to exhibit temporal trends.

Some of the alternative explanations in Table ?? can be rendered less credible by choosing a particular study design. Adjusting for known confounders and randomisation can lower the probability of confounding. Blinding can reduce the probability of performance and detection bias. Larger trials can reduce the probability of coincidence. Selecting variables \( A \) and \( B \) that do not exhibit significant temporal trends and that are spatio-temporally disjoined can reduce the probability of some of the other explanations.

In certain cases, clinical studies alone can establish that an observed correlation is causal (§??). However, establishing a causal claim in the absence of evidence of the details of the underlying mechanisms requires several independent studies of sufficient size and quality of design and implementation which consistently exhibit a sufficiently large correlation (aka ‘effect size’), so as to rule out explanations of the correlation other than causation. This situation is rare: evidence from clinical studies is typically more equivocal. Therefore, evidence of mechanisms obtained
from sources other than clinical studies can play a crucial role in deciding efficacy. Considering this other evidence is likely to lead to more reliable causal conclusions. Where this evidence needs to be considered, its quality should be evaluated in ways such as those set out in this chapter.

**Evaluating external validity**

Having established efficacy, i.e., that a causal relationship obtains in the study population, one needs to establish external validity—that the causal relationship also obtains in the target population of interest.

Recall that establishing that $A$ is a cause of $B$ requires establishing both that $A$ and $B$ are correlated and that there is some mechanism that can account for this correlation. Having established these facts in the study population, one can infer causation in the target population with some confidence if one can establish that:

1. A sufficiently similar mechanism exists in the target population, and
2. Any further mechanisms in the target population which counteract this mechanism and which are not also present in the study population do not mask the effect of the above mechanism to such an extent that a net correlation in the target population could not be explained mechanistically.

Evaluating external validity, then, requires evaluating whether the complex of relevant mechanisms in the target population is sufficiently similar to that in the study population. Evidence of mechanisms is therefore crucial to this mode of inference.

This form of inference can be especially challenging when the study population is an animal study and the target population is a human population. This is because, despite important similarities between several physiological mechanisms in certain animals and those in humans, many differences also exist. This form of inference can however also be challenging when both the study and the target population are human populations. This is because human behaviour is often a component of an intervention mechanism and may in fact hinder the effectiveness of the intervention.\(^4\)

**Other questions**

Apart from when evaluating efficacy and external validity, evidence of mechanisms can also be helpful when:

---

\(^4\) See Appendix 1.E. Well known examples involve the Tamil Nadu Integrated Nutrition Project (India) and the North Karelia Project (Finland), both discussed by Clarke et al. (2014). The Tamil Nadu Integrated Nutrition Project aimed to improve nutrition of children in school age in rural areas of Tamil Nadu and largely succeeded, while the analogous project run in Bangladesh largely failed. This was because of different social mechanisms at work in the management of the household. The North Karelia Project aimed to reduce cardiovascular disease rate in Finland by changing unhealthy behaviour. The efficacy of the project was hard to assess because both the study and control population changed their habits in response to the campaign.
• Drawing inferences about a single individual (for treatment and personalised medicine),
• Commissioning new research and devising new research funding proposals,
• Justifying the use of clinical studies, designing them, and interpreting their results (Clarke et al., 2014),
• Suggesting and analysing adverse drug effects,
• Designing drugs and new devices,
• Building economic models in order to ascertain cost effectiveness of a health intervention,
• Deciding how surrogate outcomes are related to outcomes of interest.

Example: abacavir hypersensitivity syndrome

Abacavir is a nucleoside analog reverse transcriptase inhibitor, widely used as part of combination antiretroviral therapy for HIV/AIDS, that received an FDA licence in 1998. However, its use was initially complicated by a severe, life-threatening, hypersensitivity reaction that occurred in approximately 5% of users (precise estimates vary; Clay (2002) gives a range of 2.3% to 9%). However, there was confusion regarding the cause of this reaction, and it was thought that ‘it is not possible to characterize those patients most likely to develop the HSR’ on the basis of reports of the syndrome (Clay, 2002, p.2505). This changed with the discovery that the hypersensitivity syndrome only occurred in individuals with the HLA-B*5701 allele (Mallal et al., 2002). This discovery arose from evidence of mechanisms. These authors noted that there were similarities between the mechanisms of several hypersensitivity syndromes—by ‘evidence that the pathogenesis of several similar multisystem drug hypersensitivity reactions involves MHC-restricted presentation of drug or drug metabolites, with direct binding of these non-peptide antigens to MHC molecules or haptenation to endogenous proteins before T-cell presentation.’ (Mallal et al., 2002, p.727). Patients are now genetically screened for the HLA-B*5701 allele, and this has greatly reduced the incidence of the hypersensitivity syndrome (Rauch et al., 2006).

In this document, we focus largely on the use of evidence of mechanisms to help establish efficacy and external validity. The problem of drawing inferences about a single individual is briefly discussed in Appendix 1.F.

Importance of considering evidence of mechanisms

Recall that in certain cases clinical studies on their own suffice to establish efficacy and there is no need for a detailed evaluation of other evidence of mechanisms. In other cases, however, evidence of mechanisms arising from sources other than clin-
ical studies can be decisive. In such cases, it is important to scrutinise and evaluate this evidence, just as it is important to scrutinise and evaluate clinical studies.

Situations in which it is particularly important to critically assess evidence of mechanisms arising from sources other than clinical studies include:

- Where clinical studies give conflicting results, are of limited quality, or exhibit a small effect size;
- Where randomised clinical studies are not possible, for practical or ethical reasons, in the population of interest (e.g., evaluating putative environmental causes of cancer in humans; evaluating the action of drugs in children and pregnant women);
- Where clinical studies are underpowered with respect to the outcomes of interest (e.g., when assessing adverse reactions to drug by means of studies designed to test the efficacy of the drug);
- Any question of external validity where clinical studies in the target population are limited or inconclusive;
- Assessing the effectiveness of a public health action or a social care intervention, where a thorough understanding of the relevant social mechanisms is important;
- Assessing the effectiveness of a medical device, where the mechanism of the device and its interaction with biological mechanisms may not be immediately obvious.

Some commentators have argued that one should disregard evidence of mechanisms, largely on the grounds that mechanistic reasoning has sometimes proved dangerous in the past. An infamous example concerns advice on baby sleeping position in order to prevent sudden infant death syndrome (Evans, 2002, pp.13–14). On the basis of seemingly plausible mechanistic considerations, it was recommended that babies be put to sleep on their fronts, since putting a baby to sleep on its back seemed to increase the likelihood of sudden infant death caused by choking on vomit. However, comparative clinical studies later made clear that this advice had led to tens of thousands of avoidable cot deaths (Gilbert et al., 2005). There are several other examples of harmful or ineffective interventions recommended on the basis of mechanistic reasoning (Howick, 2011, pp.154–157). As a result, it has been argued that relying on evidence of mechanisms can do more harm than good.

In many of these cases, however, the proposed evidence of mechanisms was not explicitly evaluated: often, there was little more than a psychologically compelling story about a mechanism (Clarke et al., 2014, p.350). In such cases, making the evidence explicit and explicitly evaluating that evidence would have been enormously beneficial. Thus there is a difference between mechanistic reasoning, which in some cases is based on rather little evidence and can be problematic, and evaluating mechanistic evidence, which is almost always helpful. The case of anti-arrhythmic drugs may help to illustrate this distinction. Arguably, anti-arrhythmic drugs were recommended on the basis of ill-founded mechanistic reasoning (Howick, 2011). The story goes as follows. After a heart attack, patients are at a higher risk of sudden death. Those patients are also more likely to experience arrhythmia. On the basis
of some mechanistic reasoning, it was thought likely that there was some mechanism linking arrhythmia to heart attacks. And anti-arrhythmic drugs were as a result prescribed in an attempt to indirectly prevent heart attacks by directly preventing arrhythmia. Unfortunately, it was later discovered on the basis of the Cardiac Arrhythmia Suppression Trial (CAST) that the drugs led to an increase in mortality (Echt et al., 1991). See also Furberg (1983). However, at least in retrospect, it looks as if some good mechanistic evidence had been ignored. In particular, some argued that there was little reason to think that reducing arrhythmia was a good surrogate outcome for reducing mortality due to heart attacks (Holman, 2017). In this case, properly considering the mechanistic evidence may have led to not recommending antiarrhythmic drugs.

A critic of the use of evidence of mechanisms might respond that even when there exists good evidence of mechanisms, many biomedical processes are so complex that it remains difficult to establish causal claims on the basis of evidence of mechanisms (Howick, 2011, pp.136–143). For example, there was arguably some good mechanistic evidence in favour of the claim that dalcetrapib lowers the risk of developing coronary heart disease by increasing the ratio of HDL:LDL. However, a randomized comparative clinical study showed that risk of coronary heart disease was not significantly affected (Schwartz et al., 2012). More generally, it is widely accepted that the complexity of biomedical processes presents a significant hurdle for establishing causal claims solely on the basis of evidence of mechanisms. But this is exactly why this book recommends explicitly evaluating evidence of mechanisms alongside evidence of correlation. Evidence of mechanisms is not sufficient for good clinical decision making—but neither is evidence of mere correlation.

1.2 How to consider evidence of mechanisms: an overview

This section summarises the overall approach. Subsequent sections provide a more detailed analysis.

1.2.1 Questions to address

The following protocol can be used to test a causal claim:

<table>
<thead>
<tr>
<th>Efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the effect size and quality of clinical studies establish that the observed correlation is causal?</td>
</tr>
<tr>
<td>Yes? Efficacy is established.</td>
</tr>
<tr>
<td>No?</td>
</tr>
</tbody>
</table>
• Evaluate other evidence for the claim that there exists an appropriate mechanism that can explain the observed correlation.

  – What are the hypothesised mechanisms?
  – How well confirmed is each such mechanism? What are the gaps? How well confirmed is each feature (process, entity, activity and organisational feature) of the mechanism?
  – Can the mechanism account for the full effect size? Are there counteracting mechanisms? What is the evidence that the influence of any counteracting mechanisms is less than that of the proposed mechanism?

• Evaluate other evidence to rule in or out other explanations of the correlation. Are any remaining explanations better confirmed than the hypothesis that the correlation is causal?

Efficacy is established if one can establish, in the study population, the existence of a correlation and the existence of a mechanism that can explain this correlation.

External validity
Do clinical studies directly establish a suitable association and mechanism in the target population?
Yes? Efficacy in the target population is established.
No?

• Evaluate the claim that the mechanism of action is sufficiently similar in the target and study populations.

• Evaluate the claim that in the target population, any counteracting mechanisms that are not also present in the study population do not mask the effect of the mechanism of action.

• Evaluate other evidence for a correlation in the target population.

External validity is established if one can establish similarity of relevant mechanisms in the study and target populations, and thereby establish, in the target population, the existence of a correlation and the existence of a mechanism that can explain this correlation.

In the case of efficacy, it is rare that clinical studies alone establish that the observed correlation is causal in the study population: even if the mechanism of action is not known in its entirety, there is usually some evidence of mechanisms available which contributes to establishing efficacy. Moreover, with respect to external validity it is almost never the case that clinical studies in the study population directly establish a suitable association and mechanism in the target population. Rather, external validity inferences proceed in one of the following ways (Parkkinen and Williamson, 2017):
1. By identifying and comparing the details of the mechanisms in the study and target populations.
2. Inductively, by observing a similar effect in many different experimental populations and generalizing from these to the target population.
3. Phylogenetically, by identifying the mechanism in the source population, and then inferring that the mechanisms in the source and target population are similar due to shared ancestry of the populations. The greater the degree of isolation between the target population and the source population, the less reliable this inference will be.

Thus, for both efficacy and external validity one typically needs to consider evidence of mechanisms arising from sources other than the clinical studies that establish a correlation in the study population. I.e., one needs to consider mechanistic studies.

Of course, some features of a putative mechanism may already be well established, in which case there will usually be no need to revisit the evidence for those features. Other features will be more contentious. It is only by explicitly identifying these features and the evidence that pertains to them that one can critically appraise a proposed mechanism.

1.2.2 Quality of evidence and status of claim

Quality of evidence. In what follows we shall provide guidance as to how to rank evidence for various claims: claims about correlation, claims about mechanisms and causal claims (including claims about efficacy and claims about external validity). For each kind of claim, we shall rank quality of evidence on the scale displayed in Table 1.1.

Table 1.1 Quality levels of evidence.

<table>
<thead>
<tr>
<th>Quality level</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Further research is highly unlikely to have a significant impact on our confidence in the claim.</td>
</tr>
<tr>
<td>Moderate</td>
<td>Further research is moderately unlikely to have a significant impact on our confidence in the claim.</td>
</tr>
<tr>
<td>Low</td>
<td>Further research is moderately likely to have a significant impact on our confidence in the claim.</td>
</tr>
<tr>
<td>Very low</td>
<td>Further research is highly likely to have a significant impact on our confidence in the claim.</td>
</tr>
</tbody>
</table>

Note that this system evaluates the total body of evidence pertaining to the claim in question. It does not evaluate a single study in isolation. Also, the interpretation of each category concerns the in principle possibility of obtaining further research that changes confidence in the claim. (For ethical or practical reasons, it may be
very unlikely that further research on a particular claim will be carried out; this does not imply that current evidence is high quality.)

This table of quality levels is based on the the original GRADE approach to assessing quality of evidence, put forward by Guyatt et al. (2008).\footnote{GRADE later changed the interpretation of their quality levels, dropping reference to the likelihood that further evidence will change confidence in the claim (Balshem et al., 2011, Table 2; Hultcrantz et al., 2017). This was because of concerns about the situation in which further evidence is unlikely to be obtained in practice. This change is unnecessary: as noted above, the key question is whether evidence can \textit{in principle} be obtained to significantly alter confidence in the claim. Moreover, such a change in the interpretation of the quality levels is undesirable: establishing a causal claim requires confidence in its stability as well as confidence in the claim itself. Suppose current evidence warrants 75\% confidence in a causal claim, and one learns that there is further evidence which warrants a 25\% change in confidence, but one does not know the direction of this change. I.e., one does not know whether evidence warrants 50\% confidence or 100\% confidence. Plausibly, confidence should remain at 75\% in such a situation. Arguably, however, this confidence is not sufficiently stable for the claim to be considered established or even provisionally established. This is because future evidence may be likely to decide between the 50\% and 100\% confidence, leading to a large change in confidence either way.}

\textbf{Status of claim.} In addition to the quality of the evidence, we shall also be concerned with the status that the evidence confers on the claim under consideration. The status of a claim will be measured on the scale depicted in Table 1.2.

Note that this table invokes two separate levels: the quality level applies to the total evidence, while the level of confidence applies to the claim in question. The status of the claim depends on both the quality of the evidence as well as the degree of confidence that the evidence warrants.

We will see shortly that the status of a causal claim will depend on the status of a correlation claim (assessed, e.g., by using the GRADE system) together with the status of a mechanism claim (assessed by the procedures outlined in section 4).

Appendix 1.F provides a simple probabilistic interpretation of the notion of quality and status developed in this section.

\textbf{1.2.3 Overall approach}

Figure 1.1 depicts the evidential relationships linking the concepts of this book. A claim that A is a cause of B is assessed by evaluating two further claims. The first—the correlation claim—is the claim that A and B are appropriately correlated. The second is the general mechanistic claim. In the case of efficacy, this is the claim that there exists an appropriate mechanism linking A and B that can explain B in terms of A and that can account for the extent of the correlation. There are two ways of confirming this general mechanistic claim: either via clinical studies which find a correlation that can only be explained by the general mechanistic claim being true, or by identifying the actual mechanism of action, whose features are confirmed by mechanistic studies. In the case of external validity, the general mechanistic claim is the claim that the mechanisms of action in the source and target population are suf-
## Table 1.2 Status of a claim.

<table>
<thead>
<tr>
<th>Status</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
<td>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.</td>
</tr>
<tr>
<td>Provisionally established/provisional</td>
<td>Moderate quality evidence warrants a high level of confidence in the claim.</td>
</tr>
<tr>
<td>Arguably true/arguable</td>
<td>The claim is neither established nor provisionally established, but evidence of at least moderate quality warrants significantly more confidence in the claim than in its negation, or low quality evidence warrants a high level of confidence in the claim.</td>
</tr>
<tr>
<td>Speculative</td>
<td>A claim is speculative if it falls into none of the other categories.</td>
</tr>
<tr>
<td>Arguably false</td>
<td>The claim is neither ruled out nor provisionally ruled out, but evidence of at least moderate quality warrants significantly more confidence in the negation of the claim than in claim itself, or low quality evidence warrants a high level of confidence in the negation of the claim.</td>
</tr>
<tr>
<td>Provisionally ruled out</td>
<td>Moderate quality evidence warrants a high level of confidence in the negation of the claim.</td>
</tr>
<tr>
<td>Ruled out</td>
<td>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.</td>
</tr>
</tbody>
</table>

---

**Fig. 1.1** The evidential relationships employed in this book.
ficiently similar. Again, this can be confirmed either by clinical studies on both populations that find similar correlations, or by ascertaining key features of the mechanism of action in each population and finding that these are similar. In addition, clinical studies provide good evidence of correlation, and, in certain circumstances, an established mechanism of action can also provide good evidence of correlation (see §??).

In contrast, other current EBM methods for evidence appraisal focus almost exclusively on the evaluation of clinical studies, i.e., on the two arrows at the bottom left of Figure 1.1. Moreover, they tend to conflate these two arrows—they do not distinguish the role of clinical studies in evaluating a correlation claim from their role in determining whether there is some underlying mechanism of action. Once these two roles are separated, it is clear that mechanistic studies also need to be appraised when evaluating the latter general mechanistic claim.6

Two flowcharts summarise the overall approach. Figure 1.2 depicts the workflow when evaluating efficacy. The second flowchart, Figure 1.3, applies to the evaluation of external validity. In each case there are three principal steps: gathering evidence of mechanisms; evaluating evidence of mechanisms; and using evidence of mechanisms to evaluate causal claims. Procedures for implementing the three steps are developed in §1.3, §1.4 and §1.5 respectively. The main ideas can be summarised as follows.

**Gathering evidence of mechanisms (§1.3).** It is typically more difficult to find evidence of mechanisms in the literature than it is to find relevant evidence of correlation. This is because evidence of mechanisms is characteristically produced by mechanistic studies, and there are a large number of diverse types of mechanistic studies (Smith et al., 2016). In addition, the database indexing practices for these studies tends to be unsystematic in comparison to clinical studies, for example, by lacking standardized search terms in their titles and abstracts (Evans, 2002). Arguably, this has contributed to a tendency to overlook or entirely ignore evidence of mechanisms that arises from sources other than clinical studies.

However, as explained above, such evidence of mechanisms is often crucial to establishing efficacy and external validity. Given this, the difficulties in gathering evidence of mechanisms need to be overcome. As a first step towards overcoming the difficulties, we propose a five-step strategy for identifying evidence of mechanisms, a strategy that in part relies upon existing evidence of mechanisms:

1. *Identify:* Identify a number of specific mechanism hypotheses.

---

6 The view that mechanisms should be considered alongside correlations when establishing causality is implicit in the list of causal indicators put forward by Austin Bradford Hill (1965). Several of Hill’s indicators of causality are good indicators of mechanisms, while several are good indicators of correlation. The approach of this book is very much in this tradition. We discuss the Hill indicators in relation to evaluating evidence of mechanisms in §1.4.
1. Evaluating evidence of mechanisms in medicine

Causal claim whose efficacy is to be evaluated

Do clinical studies suffice to establish or rule out efficacy?

Clinical studies and their evaluation

Formulate specific mechanism hypotheses and collect evidence (§3)

Determine the status of the claim that there is a mechanism (§4)

Determine the status of the causal claim as the minimum of the statuses of the mechanism and correlation claims (§5)

Efficacy evaluated

Fig. 1.2 Evaluating efficacy.

2. Formulate: For each specific mechanism hypothesis, formulate a number of review questions.
3. Search: Use these review questions to search the literature.
4. Refine: Identify the evidence most relevant to the mechanism hypothesis in question by refining the results of this search.
5. Present: Present the evidence relevant to the mechanism hypothesis.

This strategy is intended to help overcome some of the practical difficulties of identifying evidence of mechanisms—difficulties which may prevent practitioners from considering all the evidence. The strategy is given in more detail in §1.3.

Evaluating evidence of mechanisms (§1.4). In evaluating the quality of mechanism evidence, one should consider the following questions.

1. How well established and understood are the methods by which the evidence (of existence of a mechanism or some of its features) was produced?
   Well established methods whose functioning and potential biases are properly understood and which can be calibrated against other well established methods
typically provide higher quality evidence than methods that rely on novel techniques that cannot be calibrated against better understood methods.

2. *Can the item of evidence be produced by independent methods?*
   Employing several detection techniques and checking their results against each other is a common way to distinguish experimental artefacts from valid results. (The greater the number of independent methods that can confirm a result, the higher the quality of an item of evidence.)

3. *Are the model systems that are used in experimental research representative of humans?*
   The more faithfully model systems reproduce the relevant human features, the higher the quality of evidence gleaned from them.

4. *Can the mechanism be observed operating in many different background contexts?*
   The more robust a mechanism is against variation in background conditions, the less likely it is that our inferences based on evidence of the mechanism will err
because of unknown contextual factors interfering with the mechanism. Demonstrable robustness of the mechanism itself thus makes for higher quality evidence.

§1.4.1 and §1.4.2 describe a procedure for evaluating the quality of mechanistic studies that is broken down to three steps:

1. Evaluating methods
2. Evaluating the implementation of methods
3. Evaluating results

The status of the general mechanistic claim is then assigned as follows. A mechanism to account for efficacy is considered established when either high quality clinical studies exhibit a substantial correlation that is not explainable by, e.g., confounding or bias, or when there are high quality mechanistic studies that confirm all the crucial component features of the mechanism. A hypothesized mechanism for efficacy is considered ruled out when there is high quality evidence against the existence of the component features of the mechanism, or when high quality clinical studies consistently fail to show results one would expect if the mechanism was operating as hypothesized. A mechanism to account for external validity is considered established when high quality evidence establishes the similarity of all the crucial components of the mechanism in the study and target populations. A mechanism hypothesized to account for external validity is considered ruled out when there is high quality evidence of dissimilarity of mechanisms between the study and target populations. The more gaps or inconsistencies there are in the evidence base for a particular claim about a mechanism, the lower its status. Provisionally established claims admit some gaps in the evidence base, but require overall a good amount of high quality evidence. Arguable claims have evidence in their support that is either low quality or that has important gaps. Speculative claims are supported by evidence that show mixed results, or have little evidence in their support beyond theoretical intuition or speculation.

These issues are explained in more detail in §1.4.

Using evidence of mechanisms to evaluate causal claims (§1.5). Having ascertained the status of a correlation claim and relevant mechanism claims, one can use these to determine the status of the causal claim of interest. This process, which is explored in §1.5, may be summarised as follows.

In order to establish efficacy, one needs to establish that the putative cause and effect are correlated and that there is a mechanism that can account for this correlation. More generally, one can take the status of a causal claim to be the minimum of the status of the correlation claim and the status of the general mechanistic claim. For instance, if a correlation is arguable but the existence of any underlying mechanism is provisionally ruled out, then the causal claim itself is provisionally ruled out.

Turning to external validity, the situation is more complicated because one needs to consider (i) evidence for efficacy obtained directly on the target population, (ii) evidence for efficacy in the source population, and (iii) evidence of similarity of
mechanisms between study and target populations. Evidence directly about the target may be boosted (or undermined) by observing that efficacy does (or does not) hold in a source population that shares similar mechanisms with the target population. Table 1.4 combines the status of efficacy in the target with the status of efficacy in the study and the status of the claim that the mechanisms in target and the study are similar.

* *

Having summarised the overall approach, we now turn to a more detailed consideration of the three principal steps outlined above.

1.3 Gathering evidence of mechanisms

Suppose that clinical studies observe a correlation between the putative cause and effect, and the task is to determine whether this correlation is causal by looking for further evidence of mechanisms. In order to evaluate efficacy, it is necessary to determine the status of the general mechanistic claim, i.e., to ask whether the correlated putative cause and effect are also linked by a mechanism that can account for the extent of the observed correlation.

In the case of external validity, the existing evidence may establish causality in a study population that differs from the target population of interest. Here the relevant general mechanistic claim that needs to be evaluated is that mechanisms in the study and target population are sufficiently similar.

| General mechanistic claim for efficacy. In formulating the general mechanistic claim for efficacy, the following questions should be addressed: |
| - What is the relevant population? |
| - What is the intervention or exposure level? |
| - What is the outcome and how is it measured? |
| General mechanistic claim for external validity. In determining the general mechanistic claim concerning external validity, the following questions should be addressed: |
| - What is the target population? What is the source population? |
| - What is the intervention or exposure level in the target? |
| - What is the outcome and how is it measured in the target? |
| - What is the intervention or exposure level in the source? |
| - What is the outcome and how is it measured in the source? |

It may be that existing evidence from clinical studies together with already well-established mechanisms is enough to establish the general mechanistic claim. In
other cases, the existing evidence fails to establish causality, and it is necessary to identify and evaluate mechanistic studies. To this end, this section presents the following five-step strategy for gathering evidence of mechanisms:

1. **Identify**: Identify a number of specific mechanism hypotheses.
2. **Formulate**: For each specific mechanism hypothesis, formulate a number of review questions.
3. **Search**: Use these review questions to search the literature.
4. **Refine**: Identify the evidence most relevant to the mechanism hypothesis by refining the results of this search.
5. **Present**: Present the evidence relevant to the mechanism hypothesis.

This strategy is intended to help overcome some of the practical difficulties with identifying evidence of mechanisms—difficulties which may prevent appraisers from considering all the relevant evidence. Once this evidence of mechanisms has been identified, it can then be evaluated alongside the existing evidence of correlation from clinical studies, as explained in §1.4.

The overall approach of this section is illustrated in Figure 1.4. The five steps outlined above are explained in detail in the following subsections.

### 1.3.1 Identify specific mechanism hypotheses

**Efficacy.** In order to evaluate the general mechanistic claim that there is a mechanism that can account for the observed correlation between a putative cause and effect in a study population, it is useful to identify possible mechanisms of action. These are the specific mechanism hypotheses.

**Example:** *Specific mechanism hypotheses for determining efficacy.*

It is established that aspirin works to modify COX enzymes. The vascular benefits of aspirin understood in terms of COX enzyme generated products is established. However, ‘the mechanism of aspirin’s antineoplastic effect is less clear, with substantial evidence supporting both COX-dependent and COX-independent mechanisms. Moreover, data supporting the importance of COX-dependent mechanisms are not entirely consistent concerning the relative importance of the COX-1 and COX-2 isoforms in carcinogenesis’ (Chan et al., 2011). Chan et al conclude that: ‘Despite the large body of data regarding the potential mode of action for aspirin in chemoprevention, understanding of the mechanisms remains incomplete’. Here there is the general mechanistic claim that aspirin exhibits an antineoplastic effect. There are also a couple of more specific mechanism hypotheses, for example, that this antineoplastic effect is mediated by COX-dependent mechanisms. Evidence relating to these more
specific mechanism hypotheses provides a way to determine the status of the general mechanistic claim.

**External validity.** In order to evaluate the general mechanistic claim that *there is a mechanism in the target population sufficiently similar to the mechanism responsible for the correlation observed in the study population*, specific mechanism
hypotheses need to pertain to the mechanism of action. It is important to consider
the possibility the mechanism in the target population may contain further component
mechanisms that counteract the mechanism of action in the study population
and affect the extent of the correlation between the putative cause and effect. So one
needs to ask, are there any masking mechanisms in the target population?

Example: Specific mechanism hypotheses for determining external validity.

According to NICE guidelines, treatment for hypertension should differ depend-
ing on ethnicity (NICE, 2011). Although ACE-inhibitors have proved
beneficial for hypertension in many study populations, there remains the ques-
tion of whether they are the optimal treatment in some distinct target popu-
lation, say, African or Caribbean populations. In this case, it is necessary to
determine the status of the following general mechanistic claim: the relevant
hypertensive mechanisms in the study populations are sufficiently similar to
the mechanisms in African or Caribbean populations. This general mech-
nistic claim can be evaluated by evaluating a more specific mechanism hypo-
thesis, namely that African and Caribbean populations have a lower renin
state. As we shall see in §1.4, there is some good mechanistic evidence in
favour of this specific mechanism hypothesis, and this undermines the general
mechanistic claim. This is why instead calcium channel blockers are the rec-
ommended antihypertensive treatment in African and Caribbean populations
(Clarke et al., 2014).

There are two main ways to identify a specific mechanism hypothesis.
First, a specific mechanism hypothesis may be proposed on the basis of published
studies from the clinical study literature. If a clinical study establishes a correlation
between a putative cause and effect, and the suggestion is that this correlation is
causal, then the authors of such a study usually identify at least one possible mech-
anism hypothesis of the following form: It is plausible that mechanism X links the
putative cause and effect in the study population. The study may also point out pos-
sible masking mechanisms. Given this, the discussion section of a published paper
that reports the results of a clinical study is a good place to look in order to locate a
specific mechanism hypothesis.

Example: The discussion section of a recent paper on the effect of long-term
aspirin use on the risk of cancer says: 'Our findings suggest that for the
gastrointestinal tract, aspirin may influence additional mechanisms critical to
early tumorigenesis that may explain the stronger association of aspirin with
a lower incidence of gastrointestinal tract cancer. Such mechanisms include
modulation of cyclo-oxygenase-2, the principal enzyme that produces proin-
flammatory prostaglandins, including prostaglandin E2, which increases cel-
lular proliferation, promotes angiogenesis, and increases resistance to apopto-
sis. Aspirin may also play a role in Wnt signaling, nuclear factor B signaling, polyamine metabolism, and DNA repair’ (Cao et al., 2016). References are given for these specific mechanism hypotheses.

Second, a specific mechanism hypothesis may also be proposed on the basis of existing mechanistic studies or clinical expertise.

**Example:** It has recently been established that radiotherapy leads to a reduction in the size of large nodular goiters (Nielsen et al., 2006; Bonnema et al., 2007). But it has also long been known that there is a mechanism by which a reduction in the size of obstructions in the airway leads to an improvement in respiratory function. This was not established on the basis of clinical trials, but rather on very basic clinical experience. As a result of this experience, it may be proposed that there is a mechanism by which radiotherapy makes a positive difference to respiratory function in patients with large nodular goiters, since large nodular goiters are simply a type of obstruction in the airway that results from an enlargement of the thyroid. It may also be proposed that there is a possible masking mechanism, namely, that there is a mechanism linking radiotherapy to a swelling of the thyroid which may affect the extent of the correlation between radiotherapy and improved respiratory function (Bonnema et al., 2007).

It is important to bear in mind the following practical point. When evidence is to be evaluated by a committee of experts, it may be useful to provide a list of possible mechanism hypotheses to committee members before gathering evidence, in order to give them the opportunity to suggest alterations to the list well in advance of the committee actually meeting. Identifying a set of specific mechanism hypotheses at the outset is a good way of proceeding in the face of a large number of mechanistic studies: it makes the process of gathering evidence more manageable by helping to restrict focus to only those published mechanistic studies potentially relevant to the mechanism hypotheses of interest.

### 1.3.2 Formulate the review questions

An effective method for carrying out a review of the literature begins with a well-formulated review question. The suggestion here is to use the specific mechanism hypotheses to help formulate a number of review questions.

---

7 This routinely happens at International Agency for Research on Cancer (IARC), the UK Medicines and Healthcare Products Regulatory Agency (MHRA), the UK National Institute for Health and Care Excellence (NICE), and the EU Committee for Medicinal Products for Human Use (CHMP), for instance.
Two points are important to keep in mind:

1. Some features of the proposed mechanism may already be established, so it would be unnecessary to look for further evidence in favour of them. Such features should not figure in the more specific review questions. It is the contentious key features of the proposed mechanism that should figure in the review question.

2. The review questions may need to be updated in the course of the literature search. In particular, the search may suggest some more specific review questions about the entities, activities, and their organization in the proposed mechanism. Any changes to the review question should be documented.

Example: A number of clinical studies establish that there is a correlation between exposure to benzo[a]pyrene and lung cancer, because exposure to benzo[a]pyrene is correlated with tobacco smoking, which is itself correlated with lung cancer (IARC, 2009). But these studies alone were not be sufficient to establish causation (IARC, 2015). A number of specific mechanism hypotheses might explain the correlation between benzo[a]pyrene and cancer: e.g., (i) The diol epoxide mechanism; (ii) The radical-cation mechanism. These hypotheses lead to the following review questions concerning contentious key features of the respective mechanisms: (i) Do intermediate metabolites of benzo[a]pyrene react with DNA to form DNA adducts associated with tumorigenesis? (ii) Is benzo[a]pyrene oxidized in such a way that leads to free radical formation which may in turn form DNA adducts? These review questions can then be used to search the literature.

The review questions may be formulated according to the PICO framework. PICO stands for Population, Intervention, Comparator, and Outcome (for more information see O’Connor et al. (2011)). Consider for instance the question: Is there a mechanism in women over fifty (population) linking regularly taking aspirin (intervention) rather than not regularly taking aspirin (comparator) to asthma (outcome)?

The PICO framework helps in a number of ways to answer such a question. By making clear important aspects of the research objective it focuses the search on the most relevant literature. It also helps in the presentation of the literature that is obtained by the search. The PICO framework may be adapted to the research objective at hand. In particular, the PECO framework has been developed for non-interventional studies: Population, Exposure, Comparator, and Outcome (Vandenberg et al., 2016). One can ask, for instance: is there a mechanism in human males (population) linking exposure to high levels of benzo[a]pyrene (exposure) rather than low levels of benzo[a]pyrene (comparator) to scrotal cancer (outcome)?
1.3.3 Search the literature

A review question can then be used to search the literature for evidence for the contentious key features of a specific mechanism hypothesis. This should take place with the assistance of domain experts.

At this stage, decisions need to be made about which databases and other sources should be searched. These decisions should be documented in order to aid transparency and reproducibility. (See Appendix 1.C for some examples of databases.)

One can identify research potentially relevant to the assessment of the specific mechanism hypothesis by looking at the relevant mechanistic study literature:

1. In the first instance, this may be done by following up the references from the discussion section of any clinical study report which proposes a mechanism as the best explanation of an observed correlation. Any other publicly available reports may be useful here also, e.g., government agency reports, doctoral theses, etc.

2. More systematically, a preferred method for searching the literature may be used, e.g., a PubMed search using appropriate Medical Subject Heading (MeSH) terms, including key terms from the hypothesized mechanisms.\(^8\)

Efforts to standardise terminology and indexing practices for publications reporting mechanistic studies are welcome, especially in order to facilitate text mining techniques, which are becoming increasingly widespread. It is also important that even the negative findings of mechanistic studies are published, to reduce publication bias.

1.3.4 Refine results of the search

Identifying evidence from the literature requires expert judgement, which is susceptible to bias. In order to guard against the effects of such biases, the details of the search procedure should be clearly presented (O’Connor et al., 2011). This protects against the effects of bias by providing a transparent and reproducible literature search strategy (Vandenberg et al., 2016).

A study flow diagram can be used to present the process of selecting studies for inclusion in the review (O’Connor et al., 2011). This can be made with reference to the guidance in the PRISMA framework (Moher et al., 2009). According to this guidance, a study flow diagram consists of four phases: Identification, Screening, Eligibility, and Inclusion. After identifying studies by searching databases with a review question, the studies are then screened for duplicates, and excluded studies are recorded. The eligibility of the studies is then determined, and any ineligible studies are recorded as excluded along with the reasons for their exclusion. This leaves the included studies.

A key question here is: Is any of this evidence not relevant?

1. Use preferred inclusion and exclusion criteria and expert knowledge to rule out irrelevant mechanistic studies (Kushman et al., 2013).
   - Does the publication include original data? A good rule of thumb: if it does not include original data, then exclude the publication.

2. It may be possible to exclude some studies by a review of the title and abstract. A full-text review may be necessary to exclude other studies.
   - All excluded studies should be documented, along with the reasons for exclusion.

3. There are content management tools available to help in identifying, screening, organizing, and summarizing the evidence.
   - For example: Health Assessment Workspace Collaborative (HAWC). See: https://hawcproject.org/.

An example study flow diagram for evidence of mechanism is presented in Figure 1.5 (Vandenberg et al., 2016).

---

![Study Flow Diagram](image_url)

**Fig. 1.5** An example study flow diagram reproduced from Vandenberg et al. (2016).
1.3.5 Presenting the evidence of mechanisms

A clear summary of the identified evidence of mechanisms is an important precursor to evaluating that evidence. (Presenting the quality of evidence of mechanisms is a separate issue, for which guidance is provided in §1.4.4.) A summary of evidence of mechanisms should clearly state the general mechanistic claim that the mechanism in question is proposed to account for, that is, whether it is presented as evidence of the existence of a mechanism of action for efficacy, or as evidence of similarity of mechanisms between populations to account for external validity. This includes a clear statement of the cause \( A \) under investigation as well as the particular outcome \( B \) of interest. The presentation of evidence should also make clear the specific mechanism hypotheses under consideration, and present the evidence in favour of the contentious key features of the specific mechanism hypotheses.

**Example:** IARC’s overall process of gathering and presenting evidence of mechanisms.

In order to help identify and organize further evidence of mechanisms in the literature, the International Agency for Research on Cancer makes use of existing evidence of mechanisms in the form of ten key characteristics, one or more of which are frequently exhibited by known carcinogens (Smith et al., 2016). In our terminology, the ten key characteristic are proposals for the more specific mechanism hypotheses which are possible instantiations of the general mechanistic claim that there is a mechanism linking the considered exposure to cancer in humans. The ten key characteristics are the ability of the potential carcinogen to:

1. Act as an electrophile either directly or after metabolic activation;
2. Be genotoxic;
3. Alter DNA repair or cause genomic instability;
4. Induce epigenetic alterations;
5. Induce oxidative stress;
6. Induce chronic inflammation;
7. Be immunosuppressive;
8. Modulate receptor-mediated effects;
9. Cause immortalization;
10. Alter cell proliferation, cell death, or nutrient supply.

For instance, a correlation between benzene and cancer in humans has been observed in many studies. In order to determine whether this correlation is causal, it is necessary to determine the status of the relevant general mechanistic claim, namely, that there exists a mechanism linking exposure to benzene to cancer in humans that can account for the extent of the observed correlation (IARC, 2015). A first step is to propose specific mechanism hypotheses, with the help of the ten key characteristics. For example, the specific mechanism
hypothesis might be that benzene induces certain chromosomal aberrations that are characteristic of carcinogens. This leads to review questions that help to identify evidence relevant to this specific mechanism hypothesis. In this case, there is mechanistic evidence that exposure to benzene causes chromosomal aberrations \emph{in vivo} in bone marrow cells of mice and rats. There is also mechanistic evidence that benzene exposure also causes chromosomal aberrations and mutation in human cells \emph{in vitro}. This mechanistic evidence should be listed alongside the specific mechanism hypothesis and will adjudicate on the contentious features of the proposed mechanism. The identified evidence may be sufficient to determine the status of the general mechanistic claim, but this would involve first evaluating the evidence of mechanisms, which is the topic of the following section.

### 1.4 Evaluating evidence of mechanisms

#### 1.4.1 Overview

Evaluating evidence of mechanisms should start with clear formulations of the general mechanistic claim and each specific mechanism hypothesis, for which evidence is gathered via the procedure described in §1.3. The general mechanistic claim concerns either the existence of a mechanism (to account for efficacy) or the similarity of mechanisms between populations (to account for external validity). The specific mechanism hypotheses consider the details of mechanisms; corroborating evidence for the specific mechanism hypotheses thus supports the general mechanism claim.

Evaluating evidence of mechanisms requires assessing the reliability of the methods and techniques by which the evidence was produced. For a general mechanistic claim about the existence of a mechanism, this evidence may come from clinical studies that report a strong correlation between variables. Clinical study evidence should be evaluated according to normal criteria of good experimental design and analysis—see, e.g., Chow and Liu (2004). However, a mere correlation, even a strong one, may result from unmeasured confounding factors. Thus, only high quality clinical study evidence can significantly support a claim about the existence of a mechanism. Similarly, observing a clear dose-response relationship between variables can lend credibility to a causal interpretation (Hill, 1965), and thus to the existence of a linking mechanism. Note, however, that biological mechanisms often exhibit feedback regulation and other complex behaviours that do not give rise to clear dose-response relationships. The lack of a dose-response relationship is thus not strong evidence against the existence of a mechanism. For establishing similarity of mechanisms, one normally needs some evidence of the details of the specific features of the relevant mechanisms.
A mechanistic study provides evidence for features of specific mechanism hypotheses. Mechanistic studies are conducted by one or more of the following three means:

1. Experimental manipulation: by finding a suitable experimental system in which the mechanism or parts of it are present, making predictions about the mechanism's behaviour under interventions on some of its parts, and comparing the predictions to the outcomes of experiments where those parts are actually manipulated. Standard tools for evaluating the quality of experimental design, data analysis, randomisation procedure (when applicable) and statistical inference can thus be applied to evaluate the possibility of experimental error (Montgomery, 2009). Simulation experiments can also be used, especially to investigate whether the hypothesised organisation of a mechanism is in fact sufficient for producing the phenomenon of interest. However, the modelling assumptions on which a simulation is based should be corroborated by empirical evidence before the results of a simulation can be considered as evidence to support causal claims.

2. Observation: entities, activities and organisation of a mechanism can be found by observation techniques such as imaging technologies, autopsy, (molecular) epidemiological studies, and social surveys (for mechanisms that include parts of the social environment as components, or which are sensitive to sociological variables like socioeconomic status, parental or neighbourhood effects).

3. Analogy: Sometimes a mechanism can be hypothesised, and, to a low degree, even confirmed by analogy to an established mechanism linking a closely similar intervention/exposure to a similar outcome.

The particular challenges for evaluating evidence for features of mechanisms stem from the fact that the evidence is often produced in systems in which most of the natural context of the mechanism is absent (e.g., in vitro studies), or in which the context and possibly the mechanism itself is different from humans (e.g., model organism studies). Model organism studies are susceptible to bias in the same way as human trials. Standard ways of evaluating statistical errors or bias due to trial design may be used to assess the quality of trials conducted on experimental animals (Chow and Liu, 2004). In the case of in vitro studies that require extensive preparation of samples and that employ complicated and indirect detection methods, there is always the risk that an experimental result is an artefact produced by the instruments or preparation methods used, rather than a feature belonging to the actual mechanism. In addition to evaluating the possibility of mere experimental error and bias, weighing evidence of mechanisms requires evaluating how well these problems have been mitigated in the process of creating the evidence.

Below we describe a procedure for evaluating evidence from mechanistic studies, broken down to three steps:

1. Evaluating the methods used,
2. Evaluating the implementation of the methods, and
3. Evaluating the stability of the results.
Each step involves evaluating the mechanistic studies by means of particular quality indicators. Evidence that ranks well (respectively, badly) in the light of several indicators ought to be taken as higher (respectively, lower) quality than evidence ranks well (respectively, badly) with respect to fewer considerations. Note that this is not a rigidly algorithmic approach. Instead, domain-specific expertise should be employed in interpreting results and must be allowed to adjust the overall quality ranking. There are also trade-offs between the quality indicators; these are pointed out below. Finally, in cases where one has evidence that supports the general mechanistic claim directly, e.g. a high quality clinical trial, as well as evidence in support of some specific mechanism hypotheses (see Figure 1.1), one needs to combine these to come up with a final quality status for the general mechanistic claim.

![Fig. 1.6 A procedure for evaluating evidence of mechanisms.](image)

The procedure of this section is summarised in Figure 1.6. The three-step method for evaluating mechanistic studies in presented in the next section, §1.4.2. These steps contribute to the evaluation of the general mechanistic claim as described in §1.4.3. Finally, §1.4.4 describes how the evaluation of evidence of mechanisms can be presented.

### 1.4.2 Evaluating mechanistic studies

This section further develops the three-step procedure outlined above.

**Step 1. Evaluate methods.** The first step is to evaluate the methods employed by the studies under review. Methods should be evaluated with respect to their typical error characteristics. This requires an amount of domain specific expert knowledge, but typically there are some paradigmatic examples of well conducted studies and
reliable methods that can serve as a benchmark for evaluating the reliability of methods. A precondition for evaluating methods is that the methods themselves and their error characteristics are understood. This gives us three general quality indicators, described below.

1. **Well understood methods and model systems.** In order to evaluate mechanistic studies as high quality, it is normally essential to establish that the methods by which the evidence was produced are reliable. The better one understands how a method works, the easier it is to evaluate its reliability. Understanding how a method works is thus normally a precondition for attributing high quality to an item of evidence produced by that method. This applies to experimental model systems as well. Evidence produced in well understood model systems, in which the mechanisms responsible for the experimental result can be directly compared to relevant mechanisms in humans, should be given higher credence than evidence produced in model systems whose functioning is poorly understood. This indicator trades off against indicator (2) below: well characterised and understood experimental systems are typically simple, and thus often fail to faithfully reflect the whole-organism level physiology of humans.

2. **The degree to which experimental systems replicate human features of interest, and the quality of experimental animal trials.** Model systems that faithfully replicate human features of interest have greater external validity than ones that are very dissimilar to humans. The greater the similarity between an experimental model system and humans, the higher the quality of the evidence gleaned from the model. Notice a trade off between the choice of a model by its similarity to humans, and the tractability of the model itself. The most well understood experimental models are typically highly dissimilar to humans, whereas models that faithfully replicate many features of humans are considerably less well understood on the whole. Models that are very well characterised, but highly dissimilar to humans, are often used in basic science research that aims to discover highly general mechanisms potentially shared across many species, and such models are indispensable for this purpose. However, when the main focus of research is on justifying claims about causality in humans, the similarity of model systems to humans is an important consideration to keep in mind in evaluating evidence obtained in diverse experimental systems. This indicator trades off against indicator (1), as explained above. Studies performed on experimental animals may offer more conclusive evidence of the operation of an underlying mechanism, as more invasive intervention and measurement methods may be used in experimental animals than in humans. Animal trials are susceptible to bias in the same way as human studies, and should be evaluated similarly.

3. **The appropriateness of surrogate endpoints.** In some cases, it is not straightforward to directly measure an outcome of interest. However, it may be possible instead to measure some distinct endpoint as a way of indirectly measuring the endpoint of interest. Such a distinct endpoint is sometimes called a surrogate endpoint. For example, blood pressure may be used as a surrogate endpoint for left ventricular function, since it is more straightforward to directly measure blood pressure than left ventricular function, say, by echocardiography (Aronson,
1 Evaluating evidence of mechanisms in medicine 31

2005). Crucially, an endpoint is more likely to be an informative surrogate for the endpoint of interest if it features in the mechanism productive of that endpoint of interest. For example, there is a mechanism linking elevated cholesterol to an increase in the risk of heart disease, and so cholesterol levels are often used as a surrogate endpoint for risk of heart disease. As a result, evaluating evidence of mechanisms is important for the validation of surrogate endpoints (AHRQ, 2013). Indeed, in some cases overlooking mechanistic evidence has led to an inappropriate choice of surrogate endpoints and harmful consequences, for example, the recommendation of anti-arrhythmic drugs on the basis of employing ventricular ectopic beat as a surrogate endpoint for cardiac mortality (Holman, 2017).

**Step 2. Evaluate implementation.** The second step is to evaluate how well the individual studies have implemented the methods used. Different methods have their typical error characteristics. For instance, trials may produce biased results if randomisation is not implemented appropriately, or imaging technologies may produce artefacts. Assessing the implementation of methods consists in evaluating what means have been taken to control for the characteristic errors of the study methods. Doing this requires some knowledge of the typical error characteristics of different methods. One should thus consider the quality indicator (1) first: if the principles of operation of a particular method are poorly understood, it is more likely that one fails distinguish and control for experimental artefacts and biased results. After that, one should assess whether the methods were implemented with appropriate precautions to control for known error types. It is typically impossible to ensure that all possible sources of error have been controlled for in implementing a particular method.

**Step 3. Evaluate results.** The third step is to evaluate the stability of the results. High credence in the validity of a result can be conferred by finding that several independent methods provide similar results. This is an important indicator of the reliability of a result:

3. *Independent detectability.* The greater the number of independent methods that are able to confirm features of a mechanism, the more confident one can be that the observations are real and not artefacts.

   However, one should also assess whether results are consistent across studies conducted in similar settings using similar methods. This gives us a further quality indicator:

4. *Consistency.* Inconsistencies that cannot be explained as resulting from differences in methods or relevant contextual factors, or as resulting from poor implementation of methods in some of the studies, should result in lowering the quality status of the evidence.

   Finally, one should assess how tolerant the confirmed mechanisms are to variation in background conditions or properties of the parts of the mechanism itself.
Mechanisms that are highly robust in the sense that their operation is not disturbed by such variation are more likely to be extrapolatable between heterogeneous contexts than mechanisms that are sensitive to such variation.

5. **Robustness of features across varying contexts.** The greater the variability of contexts or model systems in which some or all features of a mechanism are found, the more plausible it is that the results are extrapolatable.⁹

### 1.4.3 Determining the status of the general mechanistic claim

This section describes how the status of the general mechanism claim can be assessed, based on the evaluation of the mechanistic study evidence for the specific mechanism hypotheses and the evaluation of the clinical study evidence for the general mechanistic claim.

Recall that different types of general mechanistic claim need to be considered for the purpose of evaluating efficacy and for the purpose of evaluating external validity. In the former case, one considers the question of whether there is a mechanism capable of accounting for the observed correlation. In the latter case, one considers the similarity of mechanisms between the study and the target populations. The two boxes below describe typical conditions in which one would attribute a high (or low) status to either type of general mechanistic claim. As evidence of mechanism can be highly heterogeneous, these conditions should not be thought of as exhaustive, nor as giving a mechanical procedure for attributing status. Instead, they are to be thought of as heuristics that need to be considered in the light of relevant domain-specific expertise, to arrive at a decision about the status of the general mechanistic claim.

<table>
<thead>
<tr>
<th>Checklist of questions to consider in evaluating a general mechanistic claim for efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Does the evidence warrant conferring a higher status to a mechanistic existence claim?</strong> Consider the following questions about the evidence; can one or more be answered in the affirmative?</td>
</tr>
<tr>
<td>1. Has a correlation of the same size been established in many studies under slightly varying circumstances (robust detectability)?</td>
</tr>
<tr>
<td>2. Is the observed correlation so large that it is very unlikely to be explained by bias or confounding, leaving the existence of a mediating mechanism as the most plausible explanation?</td>
</tr>
</tbody>
</table>

⁹ This may be understood as application of Hill’s consistency indicator to evidence of mechanisms (Hill, 1965).
3. Is the mechanism known in some detail? Can it account for the correlation and its size? Are most of the crucial features of the mechanism known and understood? Does the mechanism support novel predictions?

4. Is it plausible that the behaviour of the mechanism crucially depends on just some components or organisational features? If so, are such critical features well established according to the considerations described above? This can provide sufficient grounds for assigning the mechanistic claim a higher status than it would otherwise have. Example: consider a biochemical pathway with a single rate-limiting step. In such a case, establishing the rate-limiting step is usually more important for understanding the behaviour of the whole mechanism than establishing the rate of the reactions downstream from that step.

Does the evidence warrant conferring a lower status to a mechanistic existence claim? Consider the following questions about the evidence; can one or more be answered in the affirmative?

1. Is a counteracting mechanism likely? If so, could the correlation the mechanism is posited to explain be spurious? (If the existence of a mechanism is inferred from clinical studies, discovering that the observed correlation might be spurious counts as evidence against existence of the purported underlying mechanism as well.) If the evidence does not suggest that the correlation is spurious, this does not mean that one should revise the conclusion about the existence of a mechanism. Rather, evidence of masking suggests that the (masked) mechanism will not reliably support efficacious interventions unless the masking mechanisms can be controlled for.

2. Does the mechanism exhibit such complexity that its overall behaviour is very unpredictable?

3. Is the hypothesised mechanism inferred from evidence of an analogous mechanism in some other domain?

---

Checklist of questions to consider in evaluating a general mechanistic claim for external validity

Does the evidence warrant conferring a higher status to a mechanistic similarity claim? Consider the following questions about the evidence; can one or more be answered in the affirmative?

1. Has a correlation of the same size been established in several studies under slightly varying circumstances (robust detectability), and in several populations that are related to the target population (e.g., phylogenetically, geographically), in such a way that these correlations cannot be explained by
bias or confounding, and one must posit a similar mechanism operating in all the populations to explain the observed correlations?

2. Is the mechanism known in some detail both in the source population and the target population, and found to be similar in both, and such that it can account for the correlation observed in the source population? This can be established by applying the considerations described above.

3. When the behaviour of the whole mechanism crucially depends on some component(s) or an organisational feature, are the critical features of the mechanism similar in the study and the target populations? If so, this can provide sufficient grounds for assigning the mechanistic claim a higher status than it would otherwise have.

Does the evidence warrant conferring a lower status to a mechanistic similarity claim? Consider the following questions about the evidence; can one or more be answered in the affirmative?

1. Is a counteracting mechanism in the target population likely? Does this suggest that the correlation that the mechanism is posited to explain is spurious? If not, this does not mean that one should revise the conclusion about the existence of a mechanism. Rather, evidence of masking suggests that the (masked) mechanism will not reliably support efficacious interventions unless the masking mechanisms can be controlled for.

2. Is there dissimilarity between the mechanisms in the study and the target populations?

3. Does the mechanism proposed to support external validity exhibit such complexity that its overall behaviour is unpredictable?

4. Are the hypothesised mechanisms inferred from evidence of an analogous mechanisms in some other domain?

Mechanistic evidence for efficacy or external validity should be evaluated considering the correlational evidence that it is invoked to explain. There may be cases in which one has good evidence of mechanisms from analytical studies—e.g., from bench research on experimental systems—that could be invoked to explain a particular correlation, but the correlation in question is not itself well established. This suggests that there could be hitherto unidentified masking mechanisms that interfere with the operation of the mechanism of interest, or that the mechanism might exhibit stochastic behaviour that does not manifest as an easily detectable correlation. Such considerations should be taken into account in assessing the status of a general mechanistic claim. In evaluating a general mechanistic claim, evidence arising from clinical studies and evidence arising from mechanistic studies have mutually supporting roles.

Table 1.3 determines the status of the general mechanistic claim given the status of the general mechanistic claim based on only clinical studies and its status based on only mechanistic studies. This highlights the support the mutually supporting...
1. Evaluating evidence of mechanisms in medicine

Table 1.3 Determining the status of the general mechanistic claim (GMC) on basis of evidence from mechanistic studies and from clinical studies—points which is discussed further at the end of section §1.5.1.

<table>
<thead>
<tr>
<th>Status of GMC on basis of mechanistic studies</th>
<th>Established</th>
<th>Provisionally established</th>
<th>Arguable</th>
<th>Speculative</th>
<th>Arguably false</th>
<th>Provisionally ruled out</th>
<th>Ruled out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Provisionally established</td>
<td>Speculative</td>
</tr>
<tr>
<td>Provisionally established</td>
<td>Established</td>
<td>Provisionally established</td>
<td>Arguable</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Provisionally ruled out</td>
<td>Arguably false</td>
</tr>
<tr>
<td>Arguable</td>
<td>Established</td>
<td>Provisionally established</td>
<td>Arguable</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Provisionally ruled out</td>
<td>Ruled out</td>
</tr>
<tr>
<td>Speculative</td>
<td>Established</td>
<td>Provisionally established</td>
<td>Arguably false</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Provisionally ruled out</td>
<td>Ruled out</td>
</tr>
<tr>
<td>Arguably false</td>
<td>Established</td>
<td>Arguable</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Provisionally ruled out</td>
</tr>
<tr>
<td>Provisionally ruled out</td>
<td>Arguable</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Provisionally ruled out</td>
<td>Provisionally ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
</tr>
<tr>
<td>Ruled out</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
</tr>
</tbody>
</table>
1.4.4 Presenting the quality of evidence of mechanisms

Preparing and presenting summaries of the quality of mechanistic evidence in a standardised manner can be challenging, as evidence of mechanisms comes from highly heterogeneous sources and may involve a mixture of quantitative and qualitative relationships. Some general guidance can nonetheless be given. The following questions need to be addressed when presenting the status of the general mechanistic claim.

**Presenting the status of the general mechanistic claim for efficacy.** The following questions should be addressed:

1. What is the intervention or exposure level?
2. What is the outcome and how is it measured?
3. What is the status of the general mechanistic claim? Question to be considered here are, for instance (see §1.4.2 and §1.4.3): Does the clinical study evidence make the general mechanistic claim plausible? What are the specific mechanism hypotheses? Are there any serious gaps in the evidence for these claims? Are there any serious inconsistencies in the evidence for these claims? Is there any serious indirectness? Is counteracting plausible?

**Presenting the status of the general mechanistic claim for external validity.** The following questions should be addressed:

1. What is the target population?
2. What is the source population?
3. What is the intervention or exposure level in the target?
4. What is the outcome and how is it measured in the target?
5. What is the intervention or exposure level in the source?
6. What is the outcome and how is it measured in the source?
7. What is the status of the general mechanistic claim concerning similarity? Question to be considered here are, for instance (see §1.4.2 and §1.4.3): What is the hypothesised mechanism in the source? Are there any serious gaps in the evidence? Are there any serious inconsistencies in the evidence? Is there any serious indirectness? Is counteracting plausible? Is there any phylogenetic evidence? Is there evidence of robust efficacy in many different species?

When presenting the status of a specific mechanism hypothesis, the quality of the overall evidence of a mechanism should be presented in such a way that it also outlines the quality of the evidence for each of the individual component features of the mechanism, evaluated by employing the considerations for evaluating evidence described in section §1.4.2. For example, suppose that a drug is hypothesised
to work by binding to a particular receptor on a particular type of cell. The quality of the evidence for this interaction within the overall mechanism should be evaluated by assessing the studies providing evidence for the structure of both the drug and the receptor type, as well as any direct evidence estimating the binding affinity of the drug to its intended target. The greater the number of independent studies, employing well-established experimental methods that are able to confirm the hypothesised interaction, the higher the quality of evidence for this particular feature of the hypothesised mechanism. Conversely, if the evidence for particular features of a mechanism is inconsistent, or gleaned from few studies known to be susceptible to bias, the quality of evidence for those features of the mechanism should be considered low.

To indicate the status of particular features of the mechanism, and the general mechanism claim, one can use the following symbols:

<table>
<thead>
<tr>
<th>Status</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
<td>*</td>
</tr>
<tr>
<td>Provisionally established</td>
<td>++</td>
</tr>
<tr>
<td>Arguable</td>
<td>+</td>
</tr>
<tr>
<td>Speculative</td>
<td>?</td>
</tr>
<tr>
<td>Arguably false</td>
<td>-</td>
</tr>
<tr>
<td>Provisionally ruled out</td>
<td>--</td>
</tr>
<tr>
<td>Ruled out</td>
<td>#</td>
</tr>
</tbody>
</table>

A brief verbal explanation can be included, e.g. ++; inconsistencies. These symbols can be added to a diagram of a specific mechanism hypothesis, in order to represent the status of key features of the mechanism.

For a critical appraisal tool for mechanistic evidence which summarises key aspects of the evidence gathering process described in section 3, and the evaluation process outlined in this section, see Appendix 1.A.

This system of evaluating and summarizing evidence is not meant as a replacement for other well established evidence assessment frameworks such as GRADE. Rather, the considerations outlined here can often be integrated to existing approaches. For an example of how some of these considerations may be incorporated into the popular GRADE system by a simple amendment of the GRADE evidence profile tables, see Appendix 1.B.

**Example: ACE inhibitors.**

ACE inhibitors work by modulating the functioning of renin-angiotensin system (RAS), which is involved in regulation of the sodium concentration of blood, and arterial blood pressure. The basic architecture of RAS regarding blood pressure regulation has been corroborated by numerous studies employing varying methods—see, e.g., Fyhrquist and Saijonmaa (2008) for a review. Thus, there are no particularly contentious parts that would necessitate an in-depth evaluation of the evidence, earning the specific mechanism hypothesis
a status of established (indicated by *). This suffices to establish the general mechanistic claim in support of efficacy in those populations in which trial evidence shows a correlation between ACE inhibitor treatment and blood pressure lowering.

To establish the external validity of the blood pressure lowering effect of ACE inhibitors, one needs to establish the general mechanistic claim stating that the RAS mechanisms in the source and the target populations are similar enough. The key component of the mechanism regarding the efficacy of ACE inhibitors in African Caribbean populations is renin—an enzyme involved in the production of angiotensinogen, which is further converted by ACE into angiotensin I, and angiotensin II, a highly potent vasoconstrictor. Inhibiting ACE leads to downregulation of angiotensin II, thus inhibiting the RAS mechanism from increasing blood pressure. Low level of renin activity makes the ACE inhibitors much less effective as means to control RAS functioning. There is high quality evidence that the African Caribbean population is characterised by low renin profile (Khan and Beevers, 2005). There is thus high quality evidence that the mechanisms in white and African Caribbean populations differ at a crucial point. Thus, the general mechanistic claim that the mechanisms between these two populations are similar is ruled out (indicated by #). This is why instead calcium channel blockers are the recommended antihypertensive treatment in African Caribbean populations (Clarke et al., 2014).

Example: Evaluating dose-response relationships.

A particular challenge in evaluating the effects of a pharmacological intervention, or effects of an exposure to a chemical agent considers dose-response behaviour. Typically, dose-response is not linear, as metabolic pathways will eventually saturate as the dose increases. It may also be the case that the rate of metabolism and types of metabolites produced vary at specific doses. Normally, one does not have experimental or other data on dose-response at every level of clinical or public health interest. Rather, effects of very low or high doses must be inferred relying on models fitted to whatever data are available. This creates an extrapolation problem—how to establish that the projected responses are accurate, i.e., that the extrapolation from observed data points is reliable. Hypotheses about mechanisms often need to be considered here. For instance, assuming that dose-response is linear, and inferring hypothetical low (respectively, high) dose responses from this assumption implies that the same mechanisms, operating in the same way, are responsible for the response at all or most dose ranges. If, by contrast, measured or estimated responses suggest dose-specific effects (in the form of non-linear dose-response curve), this implies competition between dissimilar metabolic mechanisms.
An example of such an extrapolation problem comes from research on benzene. Recent evidence suggests that benzene is metabolised more rapidly at low exposures, and that low-exposure metabolism favours more hazardous metabolites (Thomas et al., 2014). If true, this implies that different mechanisms operate at low exposures than higher ones. These mechanisms should be such that they are highly sensitive to benzene—i.e., involve an high-affinity enzyme—but are quickly saturated, wherein metabolism switches to other mechanisms as the exposure increases (Rappaport et al., 2009). Estimating very low exposure levels and measuring the response can be methodologically challenging, forcing researchers to engage in extrapolations described above. Mechanistic evidence thus becomes crucial—more direct evidence of the features of enzymatic components of a metabolic mechanism that has high affinity, but gets quickly saturated, is called for. As of now, the question of low-exposure effects of benzene remains open to debate.

1.5 Using evidence of mechanisms to evaluate efficacy and external validity

In this section, we move from claims about mechanisms to causal claims, i.e., claims of efficacy and external validity. As we have seen, in order to establish efficacy, one needs to establish both the claim that there is a correlation between putative effect and putative cause and the claim that there is a mechanism connecting the putative effect and cause that can account for the size of the observed correlation. §1.5.1 explains how these two types of evidence can be combined to evaluate the status of an efficacy claim. For purposes of clinical or public health decision making one often wants to make inferences about effectiveness, i.e., about causality in target populations other than the study population. Besides evidence directly about the target population, evidence of mechanistic similarity between the target populations and study populations for which efficacy has already been evaluated may be relevant to the status of the causal claim in the target population. We deal with this question of external validity in §1.5.2.

1.5.1 Efficacy

Here we address the question of how to combine evaluations of a general mechanistic claim and a correlation claim in order to evaluate a claim of effectiveness.
General mechanistic claim. We have seen that the status of the claim that there is a mechanism connecting putative cause and effect is assessed along two different dimensions:

1. Is clinical study evidence strong enough to make it plausible that there is a mechanism that can account for the size of the correlation?
2. Is there a particular mechanism hypothesis and is the existence of the crucial features of that mechanism hypothesis established?

Correlation claim. The correlation claim is the claim that there is a correlation between the putative cause and effect, conditional on plausible confounders. Note that mechanistic evidence and results from previous clinical studies may rule in some variables as plausible confounders. Mechanistic evidence may also speak to the question whether a certain clinical study is well-conducted and properly controls for these confounding variables. Given that one has settled on both a set of potential confounders and 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. A meta-analysis, for instance, of relevant studies yields an estimate for the size of the correlation and corresponding confidence interval and $p$-value. The status of the correlation claim then depends on the width of the confidence interval, the size of the $p$-value, and the heterogeneity of the studies evaluated. A low $p$-value may, for instance, lead to a high status of the correlation claim.

Efficacy claim. To obtain the status of an efficacy claim, we combine the status of the corresponding general mechanistic claim with the status of the corresponding correlation claim. Efficacy is established just when it is established that there is a correlation and that there is some mechanism which can account for this correlation (Russo and Williamson, 2007). More generally, the status of the causal claim can be taken to be the minimum of two statuses: the status of the correlation claim and the status of the general mechanistic claim:

<table>
<thead>
<tr>
<th>Status of an efficacy claim.</th>
<th>The status of the claim $A$ is a cause of $B$ is the minimum of:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>the status of the claim that $A$ and $B$ are appropriately correlated, &amp;</td>
</tr>
<tr>
<td>2.</td>
<td>the status of the claim that there is an appropriate mechanism linking $A$ and $B$ that can account for this correlation</td>
</tr>
</tbody>
</table>

Hence, a causal claim cannot have a higher status than both the correlation claim and the general mechanistic claim. To give an example, efficacy is provisionally established if the existence of a correlation is established or provisionally established and the existence of a mechanism that can account for the correlation is provisionally established. Equally, efficacy is provisionally ruled out if correlation is provisionally ruled out and if the existence of mechanism that can account for the correlation is provisionally ruled out or of higher status.
Before turning to external validity, we discuss a potential source of confusion:

**Digression: reinforced concrete.** In the framework set out above, there are two separate distinctions in play. First, there is the distinction between evidence of correlation and evidence of mechanisms. This distinction is core to the approach taken in this handbook: the claim that \(A\) is a cause of \(B\) is evaluated according to how strongly evidence of correlation supports the claim that \(A\) and \(B\) are appropriately correlated, and how strongly evidence of mechanisms supports the general mechanistic claim that there is a mechanism linking \(A\) to \(B\) that can account for the correlation. Second, there is a distinction between clinical studies (which measure \(A\) and \(B\) together) and mechanistic studies (which investigate the details of a putative mechanism linking \(A\) and \(B\)). It is important to note that these two distinctions do not align. Both clinical and mechanistic studies can provide evidence of correlation (though clinical studies often provide better evidence of correlation than mechanistic studies). Similarly, both clinical and mechanistic studies can provide evidence of mechanisms (although mechanistic studies often provide better evidence). See Figure 1.1. Moreover, there are situations in which a causal claim can be established on the basis of clinical studies alone, as explained in §??.

Clinical studies and mechanistic studies can be mutually reinforcing. Consider an analogy to reinforced concrete, which is formed by placing steel grids into concrete. Concrete has high resistance to compressive stresses but fractures under tension. Steel, however, has high strength in tension. So, if steel is placed in concrete to produce reinforced concrete, we get a composite material where the concrete resists the compression and the steel resists the tension. The combination of two different materials produces a material that is much stronger than either of its components. In the same way, combining clinical studies with mechanistic studies produces much stronger overall evidence of efficacy than would either type of evidence on its own, because they compensate for each other’s weaknesses. For instance, clinical studies can rule out masking: masking occurs when one or more counteracting mechanisms cancel out the effect of the mechanism of action. On the other hand, mechanistic studies can rule out confounding.

The following scenarios illustrate the idea of reinforced concrete.

**Scenario 1.** Suppose, for instance, that many well conducted RCTs consistently show a correlation between the putative cause and effect and that bench research provides only very low quality evidence for the general mechanistic claim that there exists a mechanism that can account for the size of the correlation. In this case, it might seem that the correlation is established and the existence of the mechanism is speculative. In which case, efficacy is only speculative. However, this misrepresents the evidence for the general mechanistic claim. It confuses evidence obtained *only* by bench research with total evidence of mechanisms from all sources. Recall from §1.4.3 that clinical studies may also yield evidence relevant to the general mechanistic claim that there exists a mechanism. In the above example, the RCTs, when combined with the bench research, can yield a status for the general mechanistic claim that is higher than speculative—an application of the reinforced concrete metaphor. Accordingly, the efficacy claim will have a status higher than speculative.
Scenario 2. Suppose low quality clinical studies suggest that there is a correlation. Suppose too that high quality mechanistic studies support key aspects of a specific mechanism hypothesis, but that the possibility of a counteracting mechanism cannot be ruled out. In this case, it is not clear that the proposed mechanism of action can account for the observed correlation, and the general mechanistic claim will not be established. Subsequently, high quality clinical studies are carried out and determine that the net correlation is indeed positive. These studies provide evidence that any counteracting mechanism fails to totally mask the effect of the mechanism of action. The total body of evidence may now suffice to establish the general mechanistic claim (see §1.4.3). In this scenario, clinical studies reinforce mechanistic studies when evaluating the general mechanistic claim.

Scenario 3. Suppose certain clinical studies provide low quality evidence of a correlation. One might think that the key concern is confounding, so that when there is high quality evidence of mechanisms that rules out confounding, efficacy is established. However, confounding is not the only problem that arises with low quality evidence of correlation. There is also the problem that the observed correlation may not correspond to a correlation in the underlying data-generating probability distribution. In order to establish efficacy, one needs to establish that there is a genuine correlation in the underlying distribution. Hence, without high quality evidence of correlation, efficacy cannot be established.

Scenario 4. Suppose that initially, certain clinical studies provide low quality evidence of a correlation. Suppose that in this case, it is clear that the studies identify a genuine correlation conditional on certain potential confounders, but that not all plausible confounders have been controlled for. The key concern here, then, is confounding. For instance, there might be a large number of epidemiological studies all showing a correlation between putative cause and effect, but where each study fails to control for some particular variable which may be a confounding variable. Now, if there is also high quality evidence of mechanisms that rules out this variable as a confounder, efficacy is established. In this case, the mechanistic studies boost the status of the correlation claim, to established. In this case, then, the overall status is established.

1.5.2 External validity

When mechanisms within a study population and the target population are sufficiently similar, one can extrapolate an efficacy claim from the source population to the target population. In this section, we show how to combine evidence of efficacy obtained directly on the target population with evidence obtained by extrapolation from a source population.

Three assessments feed into the evaluation of effectiveness:

1. Efficacy in the target population. Although studies performed directly on the target population will normally be less conclusive than those performed on the source population, they can form the basis of a preliminary evaluation of effi-
1 Evaluating evidence of mechanisms in medicine

cacy in the target population. The preliminary status of the causal claim can be
determined as set out in §1.5.1.

2. **Efficacy in the source population.** The status of efficacy in the source population
can also be determined by considering the procedure of §1.5.1.

3. **Similarity of mechanisms in the source and target populations.** The status of
the general mechanistic claim relevant to external validity (i.e., the claim that
the mechanisms of action are sufficiently similar in source and target) can be
determined as indicated in §1.4.3.

To obtain a final status for efficacy in the target, one can combine the preliminary
status in the target population with the status of efficacy in a source population,
provided that source and target population share similar mechanisms of action. The
status of the causal claim about the target population may be increased (respectively,
decreased) by observing that efficacy does (respectively, does not) hold in a source
population that is similar to the target population. In this case, causal claims are
extrapolated from the source population to the target population.

Table 1.4 shows how the status of the causal claim in the target population can
be determined from the above three assessments. To change the preliminary status
of an efficacy claim obtained from evidence gathered on the target population, all evidence
of causation in the target population and of similarity of mechanisms needs to be
of at least moderate quality, and one or other needs to be high quality. Other quality
levels do not change the initial status.

Some remarks help to explain the table and relate it to other approaches that
address external validity.

1. If studies on the target population would on their own establish causality in the
target population, this is strong, but not infallible, evidence for causation in the
target. If there is a source population for which similarity to the target has been
established but causation has been ruled out in the source population, then cau-
sation in the target population is downgraded to provisionally established. (Note
that this situation is not covered by the protocol for evaluating external validity
advocated by the International Agency for Research on Cancer (IARC); see
Appendix 1.D for further discussion of this point.)

2. Changing the preliminary status of a causal claim obtained from evidence gath-
ered on the target population is more common when that evidence is of lower
quality. For instance, a provisionally established status may be changed to estab-
lished only in case of established efficacy in the source and established similarity
between study and target. The status arguable, however, may be changed in case
of established efficacy and provisionally established similarity.

3. The GRADE working group also evaluates whether evidence from a source pop-
ulation can be used to draw inferences about the target population. In particular,
the GRADE working group considers the case where no evidence directly ob-
tained on the target population is available.

In general, one should not rate down for population differences unless one has com-
pelling reason to think that the biology in the population of interest is so different from
that of the population tested that the magnitude of effect will differ substantially. Most
Table 1.4 Determining the status of the causal claim in the target population given the status of the causal claim in the source population, the status of the claim that the mechanism of action in source and target is similar, and the status of the causal claim in the target population on the basis only of studies carried out on the target population.

<table>
<thead>
<tr>
<th>Causation in source population</th>
<th>+ similarity of mechanism in target and source</th>
<th>Established</th>
<th>Established</th>
<th>Established</th>
<th>Established</th>
<th>Established</th>
<th>Established</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
<td>Established + provisionally established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
</tr>
<tr>
<td>Provisionally established</td>
<td>Provisionally established</td>
<td>Arguable</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
</tr>
<tr>
<td>Arguable</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
</tr>
<tr>
<td>Speculative</td>
<td>Speculative</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
</tr>
<tr>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
<td>Arguably false</td>
</tr>
<tr>
<td>Provisionally ruled out</td>
<td>Provisionally ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
</tr>
<tr>
<td>Ruled out</td>
<td>Provisionally ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
</tr>
</tbody>
</table>

often, this will not be the case. [...] The above discussion refers to different human populations, but sometimes the only evidence will be from animal studies, such as rats or primates. In general, we would rate such evidence down two levels for indirectness (Guyatt et al., 2011, p.1304-1305)

Hence, the GRADE working group takes similarity of mechanisms to be established by default when study and target populations are both human populations. This is problematic because it sets the standard of evidence required for extrapolation too low. In the case of animal studies, one can interpret the default assumption of the GRADE working group as being that the causal claim is arguable solely on the basis of causation in animals having been established. Again this is problematic. In our approach, in the absence of evidence of similarity of mechanisms, efficacy in the source population cannot be extrapolated to the target. Hence, even if many high quality RCTs in animals establish efficacy in animals, in the absence of evidence of similarity, nothing can be concluded about efficacy in humans. There is thus a sense in which the approach presented here is more cautious than the GRADE approach to external validity.

4. Causation can be established or ruled out even where no clinical studies on the target are available. This is the case when causation has been established in a source population for which it has been established that it is mechanistically similar to the target population. (This case is captured by the fourth row of Table 1.4, where causation in the target is speculative.)
1.5.3 Presenting the status of a causal claim

In presenting the status of a causal claim the following questions need to be addressed, and the status of the causal claim presented after the evaluation of evidence.

Presenting the status of the efficacy claim. The following questions should be addressed:
1. What is the population to which the status applies?
2. What is the intervention or exposure level?
3. What is the outcome and how is it measured?
4. What is the status of the correlation claim? How is this status obtained?
5. What is the status of the general mechanistic claim? How is this status obtained? (see §1.4)
6. What is the status of the efficacy claim?

The following table considers the case where efficacy is extrapolated from one to another population

Presenting the status of the effectiveness claim. The following questions should be addressed:
1. What is the target population to which the status applies?
2. What is the intervention or exposure level in the target?
3. What is the outcome and how is it measured in the target?
4. What is the source population?
5. What is the intervention or exposure level in the source?
6. What is the outcome and how is it measured in the source?
7. What is the status in the source? How is this status obtained?
8. What is the status in the target obtained by evidence directly of the target? How is this status obtained?
9. What is the status of the general mechanistic claim, i.e. that target and source are similar? (see §1.4) What is the overall status of the effectiveness claim?

Standard evidence appraisal systems can be extended to take these considerations into account. For an example of how to incorporate key aspects of this procedure into a GRADE-style evidence profile, see Appendix 1.B.
Appendices

1.A A critical appraisal tool for evidence of mechanisms

Table 1.5 below provides a critical appraisal tool for mechanistic evidence.\(^{10}\) It combines the crucial aspects of the processes of gathering evidence of mechanisms (see §1.3), evaluating it (§1.4), and using this evaluation to determine the status of a causal claim (§1.5).

Table 1.5: A critical appraisal tool for evidence of mechanisms

<table>
<thead>
<tr>
<th>What question is to be addressed with evidence of mechanisms?</th>
<th>What is best?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The main question being addressed should be clearly formulated so that one can identify mechanistic studies relevant to the causal claim of interest. The question should clearly indicate (1) the causal claim of interest, (2) whether the question concerns efficacy or external validity, (3) the general mechanistic claim relevant to the question, and (4) the specific mechanism hypotheses for which evidence is sought.</td>
<td>For guidance in formulating these questions, consult §1.3. When searching for literature, the title and abstract of a research article typically states the outcome and the proposed mechanism that is considered in a study.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Have all the important, relevant mechanistic studies been identified?</th>
<th>What is best?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The starting point for a comprehensive search for relevant studies is the major bibliographic databases, but should also include a search of reference lists that appear in relevant articles, and contact with experts, particularly to inquire about unpublished studies. The search should not be limited to the English language literature.</td>
<td>Relevant databases and experts. See §1.3 and Appendix 1.C.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What are the criteria used to select articles to be included in the evaluation?</th>
<th>What is best?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
</table>

\(^{10}\) This table is modelled on the EBM critical appraisal worksheets publicly available at the Oxford Centre for Evidence-Based Medicine website. See [http://www.cebm.net/critical-appraisal/](http://www.cebm.net/critical-appraisal/).
The inclusion and exclusion criteria should be appropriate for the question being addressed, and should be defined in advance of the search. The eligibility criteria should specify whether the studies included have demonstrated that the effect/outcome is responsive to the proposed mechanisms, and in which way. E.g., have the studies included in the evaluation demonstrated dose-dependency or not? The criteria should specify whether in vitro, in vivo, or both types of studies were included, and whether human, experimental animal, or both types of studies were included.

<table>
<thead>
<tr>
<th>Were the included studies sufficiently valid for the type of question asked?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>One should evaluate the quality of each study by using predetermined quality criteria appropriate for the type of study. Due to the nature of mechanistic studies, these considerations vary case-by-case, but typically include assessment of biological preparation and other methods used—one should consider whether the methods used can be taken to be reliable and appropriate regarding the objectives of a study.</td>
<td>The methods section of an article should describe the preparation methods, what was measured and how. Studies that do not transparently describe how the results were obtained should be excluded.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Were the results similar from study to study?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideally, the results of the different studies should be similar, but in many cases there are several specific mechanism hypotheses relevant to the outcome of interest. If the latter is the case, one should evaluate which of these mechanisms are best supported by the evidence reviewed.</td>
<td>The results section of a review article should state whether the results are heterogeneous. Similarly, the results section of individual studies should describe what the results are; when assessing many individual studies, one should compare the results across the studies to evaluate the concordance of the evidence.</td>
</tr>
</tbody>
</table>
What is the status of each specific mechanism hypothesis?

<table>
<thead>
<tr>
<th>What is best?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The review of evidence should determine the status of each specific mechanism hypothesis.</td>
<td>Considerations to take to account when evaluating the status of a mechanistic hypothesis can be found in §1.4.</td>
</tr>
</tbody>
</table>

What is the status of the general mechanistic claim?

<table>
<thead>
<tr>
<th>What is best?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Once the status of each specific mechanism hypothesis has been determined, one can evaluate the status of the general mechanistic claim.</td>
<td>A procedure for combining evidence of from mechanistic studies and evidence from clinical studies to determine the status of the general mechanistic claim is described in §1.4.3.</td>
</tr>
</tbody>
</table>

What is the status of the causal claim?

<table>
<thead>
<tr>
<th>What is best?</th>
<th>Where do I find the information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Once the status of the mechanism hypothesis has been determined, one should combine the evidence of mechanisms with evidence of correlation to arrive at a quality assignment for the causal claim of interest.</td>
<td>A procedure for combining evidence of mechanisms and correlation to evaluate the status of a causal claim is described in §1.5.</td>
</tr>
</tbody>
</table>

1.B GRADE-style tables with mechanism assessment

One widely used approach to assessing and summarizing quality of evidence and strength of recommendations in systematic reviews and clinical practice guidelines is the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system (Guyatt et al., 2011), used for example by NICE (NICE, 2014). The GRADE process involves collecting evidence to address a specific question about specific outcomes, and rating the quality of evidence according to the quality of study design, risk of bias, imprecision, inconsistency of findings, indirectness (relative to the target population), and magnitude of effect. The quality of evidence and strength of recommendation is then summarized in a table. The GRADE tables do not include an explicit assessment of mechanistic evidence. In this appendix we provide some examples of ways in which one might extend GRADE evidence profile tables to also include evidence of mechanisms. The proposed amendments are modelled according to the categories used in the GRADE tables. These amended tables illustrate that it is possible to incorporate many aspects of the approach of this book into a popular system like GRADE, without having to make any radical changes.
Table 1.6 Grade table with mechanism assessment

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Outcome 1</th>
<th>Assessment - clinical studies</th>
<th>No. of studies</th>
<th>Design</th>
<th>Risk of bias</th>
<th>Inconsistency</th>
<th>Indirectness</th>
<th>Imprecision</th>
<th>No. of participants</th>
<th>Effect</th>
<th>Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Assessment - mechanism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.6 provides a template for an augmented GRADE-style table. The proposed new categories to the GRADE table are the following:

**Mechanism hypothesis.** If the quality of trial evidence is high, and observed effect sizes sufficiently large, there may be no need to formulate and evaluate specific mechanism hypotheses. Otherwise, each specific hypothesised mechanism is sketched here.
Gaps. Crucial features of the specific mechanism hypothesis that are lacking evidence, or for which there is high risk that the available evidence is biased due to methodological limitations of the studies.

Masking. Evidence of mechanisms that counteract the effect of the hypothesized mechanism. This will reduce the plausibility of the intervention having a robust effect through the proposed mechanism.

Inconsistency. Evidence for feature(s) of a mechanism is inconsistent when there is some evidence in favour of a feature of a mechanism, and some against it, or when there is evidence for two or more mutually exclusive mechanisms. Note that inconsistency should be evaluated taking into account the amount and quality of evidence—e.g., if some of the conflicting evidence is systematically significantly less reliable due to study limitations, the inconsistency is not to be considered as severe.

Indirectness. Evidence relating to other populations and evidence of crucial differences between mechanisms in those populations and mechanisms in the target population.

In the quality and status box, one should state the overall quality of the mechanistic studies and the status of the specific mechanism hypothesis given the evidence (see §1.2.2 and §1.4). Any outstanding study limitations can be summarized here.

The overall assessment box should include an evaluation of the status of the general mechanistic claim, and should discuss how this informs the overall assessment of the status of the effectiveness claim. See §1.4.3 and §1.5.

Tables 1.7 and 1.8 depict two worked examples: assessment of brief contact interventions for reducing self-harm, and patellar taping for patellogemoral pain relief, respectively.

1.C Databases for evidence of mechanisms

Table 1.10 provides some examples of databases that can be searched for evidence of mechanisms.

1.D Assessing exposures

An important problem in causal inference in medicine involves establishing causal relationships between environmental exposures and negative health outcomes (Hill, 1965). Experimental studies, e.g., randomized controlled trials, tend to provide relatively strong evidence for causal claims. However, when assessing exposures it is typically not possible to carry out such trials in human populations, because this would involve unethically intervening to expose individuals to factors that are suspected to have deleterious health effects. The only available epidemiological studies
<table>
<thead>
<tr>
<th>Brief contact interventions</th>
<th>Repeated episode of self-harm or suicide attempt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment - clinical studies</td>
<td>12 months or less</td>
</tr>
<tr>
<td>No. of studies</td>
<td>11</td>
</tr>
<tr>
<td>Design</td>
<td>RCT</td>
</tr>
<tr>
<td>Risk of bias</td>
<td>No serious</td>
</tr>
<tr>
<td>Inconsistency</td>
<td>No serious</td>
</tr>
<tr>
<td>Indirectness</td>
<td>No serious</td>
</tr>
<tr>
<td>Imprecision</td>
<td>Some serious</td>
</tr>
<tr>
<td>No. of events</td>
<td>900/7585</td>
</tr>
<tr>
<td>Effect</td>
<td>0.87 (0.74 to 1.04)</td>
</tr>
<tr>
<td>Certainty</td>
<td>⊕ ⊕ ⊕ ⊖ Moderate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessment - mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanism hypothesis</td>
</tr>
<tr>
<td>Gaps</td>
</tr>
<tr>
<td>Masking</td>
</tr>
<tr>
<td>Inconsistency</td>
</tr>
<tr>
<td>Indirectness</td>
</tr>
<tr>
<td>Quality and status</td>
</tr>
</tbody>
</table>

| Mechanism hypothesis | Suicide prevention literacy: "BCIs improved individual's knowledge about suicidal behaviours of self-harm (e.g. risk and protective factors), what help is available, and how to access this help" |
| Gaps | No serious |
| Masking | No serious |
| Inconsistency | No serious |
| Indirectness | No serious |
| Quality and status | ⊕ ⊕ ⊕ ⊖ High, Status: Established |

| Mechanism hypothesis | Learning alternative coping mechanisms: "BCIs involved participants learning positive and functional alternative behaviours to self-harm" |
| Gaps | Some serious |
| Masking | No serious |
| Inconsistency | No serious |
| Indirectness | No serious |
| Quality and status | ⊕ ⊕ ⊕ ⊖ Moderate, Status: Provisionally established |

<table>
<thead>
<tr>
<th>Overall assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence of moderate quality shows that brief contact interventions do not appear to reduce the number of repeat episodes of self-harm or suicide attempts. Evidence of high quality shows brief contact interventions work through improving social support and suicide prevention literacy. These specific mechanisms are considered established based on good quality evidence. Evidence of moderate quality suggests brief contact interventions may also provide people with alternative coping behaviours. This specific mechanism is provisionally established. Overall, a general mechanism claim that a mechanism exists for brief contact interventions is established. However, whilst the underlying mechanisms of brief contact interventions are established, these mechanisms seem to make only a minor difference — a positive impact is not seen on repeated episodes of self-harm or suicide attempts. Further research is needed to explore alternative and counteracting mechanisms to identify mechanisms that could be targeted for more efficacious interventions, and to explain the lack of impact of brief contact interventions for self-harm and suicide.</td>
</tr>
</tbody>
</table>
**Table 1.8** Worked example 2

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Outcome</th>
<th>Assessment - clinical studies</th>
<th>No. of studies</th>
<th>Design</th>
<th>Risk of bias</th>
<th>Inconsistency</th>
<th>Indirectness</th>
<th>Imprecision</th>
<th>No. of participants</th>
<th>Effect</th>
<th>Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patellar taping</td>
<td>Immediate patellofemoral pain relief</td>
<td>6</td>
<td>4 RCTs and 2 nRS</td>
<td>Some serious risk of bias</td>
<td>Some serious inconsistency</td>
<td>No serious indirectness</td>
<td>No serious imprecision</td>
<td>-</td>
<td>166</td>
<td>-2.43 (-2.89 to -1.98)</td>
<td>⊕⊕⊖⊖</td>
</tr>
</tbody>
</table>

**Assessment - mechanism**

Mechanism hypothesis

Gaps

Masking

Inconsistency

Indirectness

Quality and status

Earlier VMO onset: "Owing to its ability to control lateral patellar tracking, vastus medialis oblique (VMO) delay or weakness is considered a key biomechanical risk factor for patellar mal-tracking. Supporting this, delayed VMO onset proved to be a risk factor for PFP development during basic military training. Additionally, a systematic review reported that VMO onset occurred after vastus lateralis (VL) in some individuals with PFP compared to controls during a range of functional tasks."

No serious gaps

No serious counteracting mechanisms

No serious inconsistency

No serious indirectness

⊕⊕⊕⊕ High. Status: Established.

Overall assessment

Evidence of low quality shows patellar taping is associated with immediate pain relief. That is, it is arguable that patellar taping provides immediate pain relief.
### Table 1.9 Worked example 3

**Intervention:** Aspirin  
**Outcome:** Cancer mortality

#### Assessment - clinical studies

<table>
<thead>
<tr>
<th>No. of studies</th>
<th>Design</th>
<th>Risk of bias</th>
<th>Inconsistency</th>
<th>Indirectness</th>
<th>Imprecision</th>
<th>No. of participants</th>
<th>Effect</th>
<th>Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Meta-analysis of RCTs</td>
<td>No serious risk of bias</td>
<td>No serious inconsistency</td>
<td>No indirectness</td>
<td>Some serious imprecision</td>
<td>98126</td>
<td>OR 0.93 (95% CI 0.84 to 1.03)</td>
<td>⊕⊕⊕⊖</td>
</tr>
<tr>
<td>51</td>
<td>IPD analysis of RCTs</td>
<td>No serious risk of bias</td>
<td>No serious inconsistency</td>
<td>No indirectness</td>
<td>Some serious imprecision</td>
<td>77549</td>
<td>OR 0.84 (95% CI 0.75 to 0.94)</td>
<td>⊕⊕⊕⊖</td>
</tr>
</tbody>
</table>

#### Assessment - mechanism

<table>
<thead>
<tr>
<th>Mechanism hypothesis</th>
<th>Gaps</th>
<th>Masking</th>
<th>Inconsistency</th>
<th>Indirectness</th>
<th>Quality and status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin downregulates COX-2 and subsequent PGE(2) formation. In addition, aspirin acetylates COX-2 to generate 'aspirin-triggered' lipoxins with anti-tumour properties. Possible other COX independent effects on oncogenic signaling. Aspirin may influence pathways critical to early tumorigenesis through modulation of a range of enzymes. These pathways are likely to impact on cancer related mortality overall.</td>
<td>Some serious gaps</td>
<td>No evidence of counteracting mechanisms</td>
<td>No serious inconsistency</td>
<td>Some serious indirectness</td>
<td>⊕⊕⊕⊖ Moderate. Status: Provisionally established.</td>
</tr>
</tbody>
</table>

**Overall assessment**

Evidence shows that aspirin use is associated with lower cancer-related mortality with moderate certainty. In other words, a correlation is provisionally established. This correlation is plausibly explained by the effect of aspirin on tumorigenesis, the general mechanistic claim for which is also provisionally established. As a result, the effectiveness of aspirin on lowering cancer mortality is provisionally established.
<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneGo MetaCore</td>
<td>Search tool for pathway analysis based on curated database of omics data, also gene-disease associations.</td>
<td><a href="https://portal.genego.com/">https://portal.genego.com/</a></td>
</tr>
<tr>
<td>GWAS catalog</td>
<td>Database of SNP-trait associations.</td>
<td><a href="http://www.ebi.ac.uk/gwas/">http://www.ebi.ac.uk/gwas/</a></td>
</tr>
<tr>
<td>Binding Database</td>
<td>Search engine for empirically verified binding affinities of drug targets, mostly proteins.</td>
<td><a href="https://www.bindingdb.org/">https://www.bindingdb.org/</a></td>
</tr>
<tr>
<td>MACiE</td>
<td>Mechanism, Annotation and Classification in Enzymes. Database of enzyme reaction mechanisms.</td>
<td>ebi.ac.uk/thornton-srv/databases/MACiE/</td>
</tr>
</tbody>
</table>

Table 1.10  A table of example databases for gathering evidence of mechanisms.

are observational. As a result, it is difficult to obtain epidemiological data that is sufficient to establish causality.

This problem occurs, for instance, when assessing whether an environmental exposure is carcinogenic in humans. In such cases, different types of evidence are needed in order to determine whether an exposure is carcinogenic to humans. For example, the International Agency for Research on Cancer (IARC) attempts to determine whether particular exposures cause cancer in humans by looking at a variety of different types of evidence, namely, epidemiological studies, studies in experimental animals, and mechanistic and other relevant data (IARC, 2015). The problem also occurs in assessing whether an exposure is an endocrine disruptor. In this context, Vandenberg et al. (2016) introduced SYRINA, a framework for the systematic review and integrated assessment of exposures. Another approach towards assessing exposures is given in molecular epidemiology. In this approach, advances in biomarker technology are utilized in order to elucidate the biological mechanisms between environmental exposures and diseases. In this appendix, we compare the approach to assessing exposures given in this book with these other prominent approaches. First we compare our approach to external validity to the approach endorsed by IARC, with reference to the example of establishing carcinogenicity of benz[a]pyrene—a compound that IARC recently evaluated and decided to upgrade from probable human carcinogen to human carcinogen largely based on
just the mechanistic evidence and evidence from cancer bioassays. Next we con-
sider molecular epidemiology as an approach that, like our approach, aims to inte-
grate mechanistic and correlational evidence to establish causality in the context of
evaluating risk from environmental exposures (Appendix 1.D.2). Finally, we com-
pare our approach to SYRINA, a framework for detecting exposures that affect the
endocrine system (Appendix 1.D.3).

1.D.1 Comparison to IARC

Here we compare our approach to external validity to that of the International
Agency for Research on Cancer (IARC). First, consider an example:

**Example:** *Carcinogenicity of benzo[a]pyrene.*

Benzo[a]pyrene is a polycyclic aromatic hydrocarbon (PAH) that is formed
during incomplete combustion of organic material. Benzo[a]pyrene and other
PAHs are found in soil, water, air, and sediments, as well as in some phar-
maceutical products, with especially high concentrations found in tobacco
smoke. Human exposure occurs through any of the aforementioned routes.
IARC has evaluated benzo[a]pyrene in four monographs, and it is currently
classified as Group 1, carcinogenic to humans (IARC, 2015).

In the most current evaluation, the IARC working group did not consider
epidemiological data, but made its decision to classify benzo[a]pyrene as car-
cinogenic to humans based on mechanistic evidence and evidence from ex-
perimental animals. This makes the case especially interesting, as according
to the procedure outlined above, the correlation between benzo[a]pyrene and
cancer required to establish the causal claim in humans would have to be in-
ferred from observed outcomes in the experimental animals together with the
mechanistic data.

On the approach of this book, first one formulates the causal claim under
scrutiny: here, ‘benzo[a]pyrene causes cancer in humans’. In the context of
IARC, this is to be taken as a qualitative claim—IARC identifies cancer haz-
ards, the exact size of the effect by which exposure increases cancer risk does
not play a role in determining carcinogenicity.

Next, one should assess—according to GRADE or a similar framework—
the evidence for a correlation between the exposure and its effect, and articu-
late any hypothetical mechanisms that would account for the correlation. The
evidence for these mechanisms should then be graded according to the pro-
cedures described in §1.4. In the latest IARC monograph on benzo[a]pyrene,
all the evidence of a correlation between the exposure and cancer came from
studies on experimental animals—no epidemiological data was evaluated. The
correlation between exposure and cancer in humans must thus be inferred
via extrapolation from corresponding data in the experimental animals. This is based on assessing the evidence for correlation in the experimental animals, and assessment of similarities of the underlying mechanisms. The IARC monograph reports high quality evidence of cancer outcomes upon exposure to benzo[a]pyrene in the experimental animals. These results are robust across eight species of experimental animals (IARC, 2009, 112–131). In addition, evidence is presented and evaluated for two main types of mechanism by which benzo[a]pyrene causes DNA adducts to form at known cancer hotspots: in one of these a metabolite of benzo[a]pyrene binding the DNA molecule, and the other an oxidized form of benzo[a]pyrene. In addition, similar activity of benzo[a]pyrene is reported to be shown in *in vitro* studies on human cell lines (IARC, 2009, 131–137).

Here IARC considered there to be sufficient evidence for carcinogenicity in the experimental animals, i.e., the causal claim about the experimental animals was established. This fits the approach of this book, as there is high quality evidence of both correlation and underlying mechanisms in the experimental animals. This alone would not suffice to transfer the same claim to humans (nor does the IARC approach consider this). However, strong evidence of similar mechanisms operating in the experimental animals and humans, and the robustness of the experimental animal results across many species, warrants a mechanism-based extrapolation of the causal claim from the experimental animals to humans. This, together with the mechanistic evidence directly on humans, such as evidence of formation of DNA adducts, is what, on the approach presented here, warrants establishing a causal conclusion about humans. In mechanism-based extrapolation, one compares the mechanisms responsible for an outcome in the target—of which a conclusion about causality is to be made—and in the source—about which direct evidence of causality is available—and looks for differences that might lead to differences in the outcome of interest between the source and the target. Here the outcome of interest is the development of tumours or the appearance of various cancer biomarkers upon exposure to benzo[a]pyrene. A dependence between these outcomes and benzo[a]pyrene has been robustly demonstrated in the experimental animals. The relevant mechanisms are the pathways by which benzo[a]pyrene causes DNA adducts that can trigger tumorigenesis, that would explain the dependence. For these, there is evidence from cultured human cell lines, as well as the experimental animals, demonstrating strong similarities, and no differences that would indicate that benzo[a]pyrene does not cause cancer in humans. In addition, there is concordant evidence of the outcomes in several species of experimental animal, lending further credibility to the assumption that the carcinogenicity of benzo[a]pyrene is not dependent on idiosyncratic features of any particular species. These considerations, taken together, suffice to establish the carcinogenicity of benzo[a]pyrene in humans.
While the approach of this book would yield the same conclusion as IARC’s, it should be noted that the procedures differ at certain points. IARC does not formally endorse extrapolation from experimental animals, nor robustness of evidence as grounds for upgrading a classification, but allows for upgrading (or downgrading) a classification of carcinogenicity on the basis of mechanistic evidence alone. On the approach of this book, one may appeal to the aforementioned considerations, and one needs in addition to establish correlation in humans (by direct observation or extrapolation), before any claim about causality can be considered established.

Having considered an example, we now compare the general approach of this book to external validity to that of IARC. IARC’s approach is summarized in Fig. 1.7.

![EVIDENCE IN EXPERIMENTAL ANIMALS](http://monographs.iarc.fr/ENG/Publications/Evaluations.pdf)

**Fig. 1.7** IARC’s approach to classifying potential carcinogens (http://monographs.iarc.fr/ENG/Publications/Evaluations.pdf).

The categories of IARC roughly correspond to those presented here, as follows. IARC have a ranking for overall carcinogenicity:

- **Group 1**: Established
- **Group 2a**: Provisionally established
- **Group 2b**: Arguably true
- **Group 3**: Speculative
- **Group 4**: Ruled out
IARC also has a separate ranking of evidence of carcinogenicity in humans and animals:

- **Sufficient:** Established
- **Limited:** Provisionally established
- **Inadequate:** Arguable or speculative

**Evidence Suggesting Lack of Carcinogenicity (ESLC):** Ruled out

The approach of this book is simpler than that of IARC in one respect: a single scale from established to ruled out, rather than two different categorisations. On the other hand, the scale adopted here involves more categories.

Moreover, this book draws a distinction between establishing causality in the source population and establishing similarity of mechanism; these differences mean that there is no exact analogue of the above table for the approach of this book.

**Table 1.11** Determining the status of the causal claim from similarity of mechanisms in the source and target populations and causation in the target population on the basis of evidence obtained on the target population. It is assumed here that causality in the source population has been established.

<table>
<thead>
<tr>
<th>Similarity of mechanism in target and source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
</tr>
<tr>
<td>Established</td>
</tr>
<tr>
<td>Provisionally established</td>
</tr>
<tr>
<td>Arguable</td>
</tr>
<tr>
<td>Speculative</td>
</tr>
<tr>
<td>Ruled out</td>
</tr>
</tbody>
</table>

In order to compare the approach of this book with that of IARC, consider two tables that illustrate the approach that this book takes with respect to external validity. First, Table 1.11 assumes that causality in the source has been established and charts similarity of mechanisms in the source and target populations against causation in the target population on the basis of evidence obtained on the target population. A second table, Table 1.12, assumes that similarity of mechanism is established and charts causation in the source population against causation in the target population on the basis of evidence obtained in the target population.

There is a broad agreement between the approach presented here and that of IARC. As with the approach advocated here, IARC employs evidence of mechanisms to draw conclusions about causation at two places: to evaluate efficacy in humans on basis of evidence directly in humans and to ensure that causal claims in specific animal populations can be extrapolated to humans. For the first task, IARC employs the Hill criteria without assessing mechanistic studies in a systematic way. It is only in assessing external validity that IARC explicitly evaluates studies that investigate the details of the mechanism of action.
Table 1.12 Determining the status of the causal claim from causation in the source population and causation in the target population on the basis of evidence obtained on the target population. It is assumed here that similarity of mechanism has been established.

<table>
<thead>
<tr>
<th>Causation from target studies</th>
<th>Established</th>
<th>Provisionally established</th>
<th>Arguable</th>
<th>Speculative</th>
<th>Ruled out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Provisionally established</td>
</tr>
<tr>
<td>Provisionally established</td>
<td>Established</td>
<td>Provisionally established</td>
<td>Arguable</td>
<td>Arguable</td>
<td>Speculative</td>
</tr>
<tr>
<td>Arguable</td>
<td>Established</td>
<td>Provisionally established</td>
<td>Arguable</td>
<td>Arguable</td>
<td>Speculative</td>
</tr>
<tr>
<td>Speculative</td>
<td>Established</td>
<td>Arguable</td>
<td>Speculative</td>
<td>Speculative</td>
<td>Ruled out</td>
</tr>
<tr>
<td>Ruled out</td>
<td>Provisionally ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
<td>Ruled out</td>
</tr>
</tbody>
</table>

The approach presented here is more explicit with respect to where and what evidence of mechanisms should be used. Firstly, this book recommends explicitly evaluating mechanistic studies when evaluating evidence obtained directly in humans. After all, evaluating both whether there exists a mechanism and whether there exists a correlation is necessary for evaluating the evidence obtained directly in humans (§1.5.1). The Hill criteria can only be seen as a first approximation to the comprehensive assessment of mechanistic evidence needed to establish efficacy in humans. What is more, these criteria obfuscate an otherwise clear distinction between evidence pertinent to the correlative claim and evidence pertinent to the general mechanistic hypothesis (§1.4).

Secondly, this book separates the overall evaluation of causality and the evaluation of evidence directly obtained in humans. The overall evaluation is obtained by aggregating the evidence directly obtained in humans and the evidence in animals (§1.5.2). For instance, it might be that, initially, efficacy is established in humans by considering studies that purely involve humans, but that, subsequently, studies of a variety of animal species that are mechanistically similar to humans rule out causation in those species. These further studies would surely cast enough doubt on causation in humans so that the causal claim can no longer be considered established. However, by identifying the overall evaluation with the evaluation of evidence directly obtained in humans when the evidence obtained on humans is sufficient (see the top row of the IARC table, Figure 1.7), IARC assigns Group 1 in this case (the top right-hand corner of the IARC table). The procedure set out in this book would assign status established to the efficacy claim on basis of just the evidence directly obtained in humans, but it would assign overall status provisionally established on the basis of all the evidence, animal as well as human (see the top-right corner of Table 1.12). This classification is arguably more appropriate.
1.D.2 Molecular epidemiology

A key working hypothesis in epidemiology is that diseases are often the result of environmental exposures. However, despite much epidemiological research, the biological processes linking many environmental exposures and diseases remain unknown. This is the case despite the technological advances in measuring certain biomarkers, i.e., biological markers of events at the molecular and physiological levels. Unfortunately, it is precisely the details of these biological mechanisms that are required in order to establish efficacy and external validity in the absence of sufficient clinical studies.

Molecular epidemiology is a response to this state of affairs (Schulte and Perera, 1993). Molecular epidemiology is a branch of epidemiology that makes use of advances in biomarker technology in order to elucidate the biological mechanisms between environmental exposures and diseases. An important methodology in molecular epidemiology involves utilizing complementary studies to validate biomarkers that mediate between environmental exposures and disease outcomes (Vineis and Perera, 2007). For example, some studies may provide information correlating a certain biomarker to a particular environmental exposure. Other studies may provide information correlating a disease outcome to the same biomarker. By bringing together the results of these studies, the disease may be correlated with the environmental exposure while at the same time providing some insight into the biological mechanisms responsible for this association by highlighting the intermediate biomarkers (?).

In the case of benzene and leukemia, studies revealed that certain chromosome aberrations were predictive of cancer in humans (Bonassi et al., 2000). In other case-control studies, those same chromosome aberrations were seen to be more frequently present in leukemia patients that had been exposed to benzene (Zhang et al., 2007). Not only, then, was a chain of correlations established between benzene and leukemia but also some insight was provided into the biological mechanism underlying this chain, viz., the role of chromosomal aberrations (Vineis and Perera, 2007). This approach in molecular epidemiology thus complements the approach advocated in this book. For more information, see Wilde and Williamson (2016).

1.D.3 Comparison to SYRINA

SYRINA (Vandenberg et al., 2016) is a framework that was put forward to evaluate the strength of evidence that a certain exposure is an endocrine disruptor. This approach first evaluates the evidence for an association between chemical exposure and (adverse) effect. Second, this approach evaluates the evidence for an association between the chemical and endocrine disrupting activity. Third, the evidence for an association with an (adverse) effect and for an endocrine disrupting activity are combined to obtain an overall assessment of endocrine disruption.
SYRINA combines quality of evidence ratings from different streams of evidence in all three steps. As with our approach, the quality level of the causal claim is the minimum of the quality of the different evidence streams. Fig. 1.8 gives the relevant SYRINA table for an association between chemical exposure and (adverse) effect.

![SYRINA table](image)

**Fig. 1.8** SYRINA table for an association between chemical exposure and (adverse) effect.

The resulting initial rating can be upgraded by one level if there is high confidence in the evidence from in silico and in vitro studies.

In the next step, the endocrine disrupting activity of the exposure is evaluated by combining different evidence streams. This time in vivo and in vitro evidence is combined. Fig. 1.9 gives the relevant SYRINA table.

Finally, the quality levels for the association with adverse health outcomes and for the endocrine activity are combined according to the table in Fig. 1.10.

In relatively unusual cases the resulting quality level can be upgraded or downgraded by considerations given to the plausibility of the link of disrupting endocrine disrupting activity and outcome.

Let us consider some points of comparison between SYRINA and the approach of this book. First, this book formulates explicit methods for evaluating evidence of mechanisms (§1.4). Second, for the evaluation of both endocrine activity and association with adverse health outcomes, SYRINA only combines two kinds of study. When evaluating the plausibility of an association with adverse outcomes, SYRINA combines results from experimental laboratory animals with evidence in humans or wildlife animals. According to the approach presented in this book, application of results from such associations in animals would need to be extrapolated with the help of evidence of mechanisms along the lines of §1.5.2. In addition, mechanistic considerations may be relevant when evaluating whether there is an association of the chemical with adverse health outcomes. After all, an observed correlation may be due to confounding. As with IARC, SYRINA makes use of the Hill Criteria for
evaluating each stream of evidence and does not explicitly distinguish between evidence of mechanism and evidence of correlation. Hence, while this book agrees with SYRINA that many evidence streams should be considered when evaluating causal claims, we would emphasise the need for a more systematic integration of evidence of mechanisms and evidence of correlation along the lines of §1.4 and §1.5.

1.E Assessing mechanisms in public health

Introduction. When applying the ideas described in this book to areas other than therapeutic clinical medicine, a number of further considerations need to be borne in
mind. The arena outside of clinical medicine where most thinking has been done relating to methods of evidence appraisal is public health (NICE, 2012). Public health is concerned with the actions, interventions and policies designed to protect the public from hazards, to prevent disease, and to promote good health. In different countries, specific institutions were given the task of developing methods for the assessment of evidence and for the formulation of guidelines in public health. These individual efforts have been brought together into a European initiative, led by the European Centre for Disease Prevention and Control (ECDC). In their 2011 synthesis report, they show how public health should adopt and integrate the methods of evidence-based medicine, specifically the GRADE system, for the assessment of evidence.11 In this appendix the focus is on one particular sub-issue, namely mechanisms of causation and, given the concerns of this book, how to deal with mechanisms conceptually and then practically in the appraisal of evidence.

Public health and evidence-based medicine in the UK. Public health in the UK has been working within the evidence-based paradigm formally since 2000, and much has been learned (Kelly et al., 2010; Kelly and Moore, 2012). In 2001 the English Department of Health published its Research and Development Strategy. Amongst other things it made the case for using the principles of evidence based medicine in public health (Department of Health, 2001). Organisations such as the Centre for Reviews and Dissemination at the University of York, the Cochrane Collaboration, the Campbell Collaboration, the Health Development Agency and NICE took up the challenge. These organisations have confronted in various ways the methodological, theoretical, practical, epistemological and ontological problems of applying EBM principles to the very broad church of public health. Since then other policy areas have gone in the same direction of taking an evidence based approach. So social care, education and criminal justice, amongst others, have all had agencies created to move these arenas onto an evidence based footing (Paisley et al., 2017).

Statistical associations and correlations in public health. Statistical associations and correlations have been at the heart of progress in public health for many years. A number of landmark studies show just how important finding statistical associations can be. The investigations by Doll and Hill (1950, 1952, 1964) into the connections between smoking and disease are the original benchmarks. Their initial observations showed that there was an association between exposure to cigarette smoke and carcinoma of the lung (an association which had not hitherto been noticed). This led, in the long run, to public health policies which have reduced the prevalence of cigarette smoking in the population and greatly reduced the number of deaths from lung cancer, and also heart disease, stroke, and various other cancers, which were subsequently found to be associated with exposure to cigarette smoke too.

These pioneering works are often thought to be purely statistical, but in fact Hill was concerned with biological plausibility, and hence mechanisms (Hill, 1965). Since the early 1950s when the first statistical observations were made, the biological mechanisms operating in the interaction between the contents of cigarette smoke and the tissues in the lung, as well as the mechanisms relating to the effects on blood circulation, heart functioning, arterial disease, and many other pathologies have been described. Considerations about biological plausibility led, for instance, to investigations of the relation between asbestos and mesothelioma (Doll, 1955; Newhouse and Thompson, 1965). Scientific discoveries relating to these mechanisms continues to the present. The basic mechanisms are well understood in individual human beings, and public health policy has developed in such a way that smoking in the European Union is now a minority habit and protection from unwanted exposure to cigarette smoke is the norm.

So cigarette smoking was identified as what public health practitioners have come to call a risk factor. In the wake of this great public health success, statistical associations have emerged over the years pointing to risks from other things, notably a lack of physical activity, being overweight and obese, over consumption of alcohol (Sytkowski et al., 1996), certain types of sexual activity (Dougan et al., 2005), ingesting certain non-prescribed drugs (White and Pitts, 1998) as well as toxins in the environment, although the dangerous consequences of exposure to certain substances used in industrial processes like asbestos, phosphorus and radium had been known long before the discoveries about smoking (Gochfeld, 2005).

There is now a very large and important scientific literature originating in the observation of statistical correlations and subsequently strengthened into causal understandings based on the mechanisms at work in the human body following exposures. Policies designed to protect the public have flowed from this scientific knowledge. New risks regularly appear and currently the role of air pollution and toxins from emissions from vehicles is under scrutiny. This debate mirrors events in the 1950s when the dangers from smog in urban environments caused by the burning of coal led to the Clean Air Act and the phasing out of coal as a primary domestic fuel in the UK (Brimblecombe, 2006). In public health there is a long history of bringing together correlations and mechanisms to understand the processes which can cause a number of very common diseases and which potentially offer a platform to take action to mitigate the risks and harms, and, as with the Clean Air Acts of the 1950s and action against tobacco, have been highly effective and successful.

**Recurrent public health problems—non-communicable disease in the present.**

However, notwithstanding the successes with smoking and clean air, deaths from preventable causes which are known and well understood have not gone away. Deaths from non-communicable diseases associated with excess calorie and alcohol consumption and lack of physical activity continue to increase steadily in most countries around the world (Beaglehole et al., 2012). Type 2 diabetes, cardiovascular disease, and certain cancers all have rising prevalence even though the statistical associations between the diseases and the risk factors are well known and the mecha-
Evaluating evidence of mechanisms in medicine

1 Evaluating evidence of mechanisms in medicine 65

nisms operating at the individual level are well understood (though in some diseases better than others).

This is fundamentally important as far as appraising evidence of mechanisms is concerned. It is fundamentally important in ethical terms too, because the rising prevalence, while affecting the whole of the population, affects those in poorer and more disadvantaged circumstances to a far greater extent than the well to do and the privileged (Wilkinson and Marmot, 2003). There is a sharp gradient in health inequities that shows a strong correlation between poor health and early death from non-communicable disease and disadvantage. This holds whether disadvantage is measured by income, occupation (or lack of it), housing tenure or educational level or qualifications (Buck and Frosini, 2012). The fact is that there are a number of mechanisms which are conceptually and practically distinct from the mechanisms describing the processes of the of disease causation following exposure to a pathogen or toxin of some kind. Such mechanisms operate at the group and individual level, and concern social interactions and support, access to socio-sanitary infrastructures, psychological factors, etc. It is these mechanisms as well as the biological ones, which have to be explored in the appraisal of public health evidence (Kelly et al., 2014).

The individual level and the population level. The first thing to note is that mechanisms operate at different levels. In almost all of the investigations referred to above, the mechanisms that have been subject to most scrutiny are those operating at the level of individual human biology. So, after association were found in the population data, the focus shifted to understanding what was actually going on in the human body when it was exposed to cigarette smoke, ethanol, high levels of sugar, asbestos, particulates in the atmosphere and so on. And this approach of course has shown why these exposures are harmful and how they operate on the human biology. These investigations have been extremely successful and we now have plausible biological mechanistic explanations.

But what about the mechanisms operating at the population level? What about the mechanisms that produce the patterning of health between the rich and the poor, between different parts of countries (Graham and Kelly, 2004)? In the United Kingdom, for instance, health on average is much worse in Scotland and the North of England than in the South. How can we explain that? What are the mechanisms which explain the fact that, on average, baby boys born in Guildford will live fifteen years longer than baby boys born in Shettleston in Glasgow? What are the mechanisms which link poverty to early death? And what are the relationships between the mechanisms going on biologically and in the wider social and physical environment (Kelly et al., 2014)?

With the stunning progress in understanding the biochemistry of disease since the nineteenth century, the tendency has been to first focus on mechanisms operating at the biological individual level. As noted above this is usually relatively straightforward, as the biological processes have been well understood in broad terms for decades and the detail is constantly developing as the science progresses. But what are the social and behavioural mechanisms involved? The behavioral mechanisms
are also reasonably well described in the psychological literature (see Table 1.13 for some examples). Models and theories explaining why, on average, humans are likely to do this or that, are plentiful (Conner and Norman, 2005). However, why when following the same intervention based on the same information about the dangers of smoking, one individual does "this" (say, decides and successfully quits smoking) and one does "that" (doesn’t even think about quitting smoking) is less well understood in a mechanistic sense (Marteau et al., 2015).

<table>
<thead>
<tr>
<th>Hazard</th>
<th>Behavioural mechanisms</th>
<th>Disease/morbidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure to cigarette smoking</td>
<td>Teenagers imitating peers or smoking parents</td>
<td>Lung cancer, cardiovascular disease</td>
</tr>
<tr>
<td>Exposure to ethanol, binge drinking</td>
<td>Socializing</td>
<td>Liver disease, certain cancers</td>
</tr>
<tr>
<td>Exposure to HPV</td>
<td>Unprotected sex, socializing</td>
<td>Cervical cancer</td>
</tr>
<tr>
<td>Exposure to mosquitoes</td>
<td>Sanitation, clothing</td>
<td>Zika</td>
</tr>
<tr>
<td>Workplace posture</td>
<td>Incorrect posture while sitting or working with a computer, poor lifting practices</td>
<td>Lower back pain</td>
</tr>
<tr>
<td>Work overload</td>
<td>Organizational structures, management practices</td>
<td>Anxiety, depression</td>
</tr>
</tbody>
</table>

However, where the biggest gaps in mechanistic understandings exist, is at the social or population level. The associations between poverty and poor health have been known since at least the middle of the nineteenth century and for probably much longer than that in a non-statistical sense. But how it works mechanistically is much less well defined. From an evidence appraiser’s point of view there is no easy solution to these problems and neither will there be till primary studies examining the mechanisms have been conducted. But it is important nevertheless to ask the questions. And to ask the questions in a way that acknowledges that we do indeed know with a very high degree of certainty that there is a relationship between wealth, education and employment and health, but we do not know with sufficient clarity what the mechanisms are and in such a way as to target interventions and policies in a directed way to be maximally effective (Kriznik et al., 2017).

There have been many attempts around the world to tackle inequalities in health and while overall the health of populations has improved decade on decade, the relative inequities remain a stubborn fact of life (WHO, 2008). Although the lack of political will to do something about it has been a major barrier everywhere, one of the other important reasons for failure has been an absence of mechanistic studies at the population level studies and therefore the inability to know what to do based on mechanistic understandings of the causal pathways involved.

**The biological level and the social level.** In recent years, the relationship between the individual biological level and the social level has come under scrutiny as a consequence of developments in biology itself, particularly developments in developmental programming, epigenetics and metabolomics. While each of these
topics is different, what they have in common is that they show how the human phenotype is the product as much of its environment physically and socially as it is of its genetic inheritance (Kelly and Kelly, 2017). Human (and animal) biology is much more plastic in the face of environmental exposures than had been previously thought. DNA doesn’t change, but the way that it is expressed does. The metabolite structure of our bodies reveals a timeline of the various exposures we have been subjected to across the life course. Factors affecting the health of our grandmother when she was pregnant with our own mother may have a fundamental effect on our own health in adulthood. The mechanisms here are now quite well developed (Hanson and Gluckman, 2011; Ozanne and Constância, 2007) and they show that our health is not just a metabolic response to toxins; it is about a complex social and biological interaction—a relational process or mechanism. These mechanisms are critically mediated by the social worlds that people inhabit.

This science is still developing at a rapid rate and along with it, the understanding of the human genome and the therefore of individual biological differences between humans. It is highly likely that new and better mechanistic models and understandings will emerge including ones incorporating the social factors. The implications for the evidence appraiser at this stage are that the question should be asked—are mechanisms relating to the relationship between biological and social factors being described, used, and articulated? A further important epistemological consideration is the degree to which the approach taken by the researcher is a genuinely a relational one—in other words, one that sees the process as a dynamic and interactive one rather than a deterministic one. This is important because if the new understandings of the plasticity of biology are to be useful in public health, the models need to move away from a reductionist approach and should instead be about elucidating the interactive nature of the process. Again this is a question to be asked by the evidence appraiser: what is the nature of the interaction?

Mechanisms of disease and mechanisms of prevention. There is another question to be asked about the evidence of mechanisms in public health matters and that is about the difference between the causes of disease and the causes of prevention (Kelly and Russo, 2018). So far in this appendix we have focused on the important difference between the causes of disease in individuals and the mechanisms involved and the causes of the patterning of disease at population level and the mechanisms involved in this patterning. We have also discussed the mechanisms involved in the relationships between the two.

But there is another very important distinction to draw out which is especially important in public health. This is the difference between mechanisms causing the disease (either in individuals or in populations) and the mechanisms involved in preventing disease (e.g., Table 1.14). The question simply is this. Does knowing the cause of a disease (an exposure to something which is risky) and knowing that by reducing exposure that disease will be prevented, tell you how to reduce exposure? The short answer is that it doesn’t, though many public health policies proceed as if it did. The biology of the aetiology of lung cancer, of liver disease, of type two diabetes and the metabolic syndrome tell you nothing about the mechanisms involved
in helping people to stop smoking, to consume less alcohol, to eat fewer calories or take more exercise. Knowledge of the cause tells us what people should do, but it doesn’t explain how to do it. The mechanisms involved in smoking and giving up smoking, the mechanisms involved in the practices of eating and drinking (and for that matter, sexual conduct, bad driving, or going jogging) belong to a quite different realm of evidence than microbiology. The relevant evidence is social and psychological. The mechanisms involved are social and psychological and there is a considerable amount of evidence, some of which has been around for a long time, describing both associations and mechanisms—see Becker et al. (1977) and Kelly and Russo (2018). For the most part, however, public health policy (with the very significant and successful exception of smoking) pays scant attention to the social and psychological evidence, mechanistic or otherwise. We suggest that the evidence appraiser begins by asking the question: what evidence is available about the aetiology of the disease? And what evidence about effective preventive measures? The distinction between aetiology and prevention should then guide the appraisal of correlations and of mechanisms. Specifically, are only mechanisms at biological level invoked, or also social mechanisms?

Table 1.14 Public health mechanisms for tackling obesity

<table>
<thead>
<tr>
<th>Generic, population level</th>
<th>Targeted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food advertising, e.g., max amount per day, recommended amount per day, amount of lipids or carbohydrates contained in food portions</td>
<td>MEND programmes in the UK <a href="http://www.mendfoundation.org/">http://www.mendfoundation.org/</a>, e.g., targeted training for school children about diet and healthy lifestyles, targeted training for parents about psychological risks related to obesity</td>
</tr>
</tbody>
</table>

Finally, for both mechanism of disease and mechanism of prevention, the evidence sources will be heterogeneous. The disciplines of psychology, sociology, economics, anthropology, organisational behaviour, political science, history, and the public health sciences all have, and have had, things to say on these matters. Unfortunately, it is not the case that we can simply cheerfully agree that the evidence for these things is heterogeneous so we should just pull it all together, synthesise it and out will come a nice clear set of mechanisms. The reason for this is that each of these disciplines, and the many sub-disciplines within each of them, operate with a variety of epistemological, methodological and ontological assumptions about the nature of human life and its place in the world. Sometimes these veer toward highly individualistic accounts sometimes to more socially oriented accounts. So the task is not to try to adjudicate, but to acknowledge the differences, to articulate them (even if the researchers don’t themselves do that), and to consider the degree to which the different positions really matter in terms of the substantive problem (Kelly, 2017). Intriguingly, all these disciplines are dealing with the same basic concern—humans in the physical and social world and what is going on in their heads as they go about their business. They each construct ways of seeing and describing the same phenomena differently and in ways that sometimes defy any kind of commensura-
bility. However as long as the appraiser keeps in mind that the basic thing under consideration is the same, and there are just lots of different ways of looking at the phenomena, then the task is not an impossible one. But as ever the first step is to ask the appropriate question, to describe what is there in terms of evidence and to determine to what extent this allows us to understand the mechanisms with clarity.

Here are some simple questions that one can ask in order to structure the search for relevant mechanistic studies, in the context of public health interventions:

**Checklist of questions:**

1. What disease is the intervention targeting? Infectious or non-communicable?
2. What biological mechanisms are known?
3. What socio-economic or psychological mechanisms are known scientifically?
4. How can behavioural mechanisms reduce exposure?
5. Why might public health interventions targeting the pathogens fail?
6. What is the public perception of the disease in terms of risk, seriousness and personal vulnerability?
7. What mechanisms come into play as a population or different segments in the population react to an intervention?
8. Are there sub-groups within the population that should be specifically targeted? How can they be reached and what specific mechanisms might come into play?

### 1.F Particularisation to an individual

Inference from an effectiveness claim involving a whole population to effectiveness in one of its members is of central importance in medical diagnosis, prognosis, and treatment. This mode of inference is often called *direct inference*.

If one has established effectiveness in a population, then one has also established that there is a mechanism operating that connects the putative cause and effect. Now, the population may not be entirely homogeneous with respect to this mechanism: some individuals will exemplify the mechanism while others may not. One way to establish that mechanisms in the population are applicable to a particular individual is by assessing how homogeneous the population is with respect to the mechanism of action. Inference from a homogeneous population to individuals is more likely to succeed, because most individuals will exhibit the mechanism responsible for causation in the population.

However, in most cases there will be subpopulations for which effectiveness does not hold. There may be several reasons for this kind of *exceptionality*. Firstly, in some such subpopulations the mechanism responsible for effectiveness in the
whole population simply does not operate. For instance, while drinking considerable amounts of milk is normally safe, subpopulations with lactase deficiency should drink only small amounts of milk. Considering whether crucial features of the mechanism responsible for effectiveness are present in the particular individual can therefore increase certainty about whether the causal claim is applicable to the individual. Secondly, counteracting mechanisms may operate in some subpopulations. For instance, exercising is normally beneficial for preventing stroke by lowering blood cholesterol, but smoking may counteract these beneficial effects by raising blood cholesterol. With this in mind, the following questions can assist the evaluation of evidence of mechanisms for direct inference:

**Particularisation to an individual**

What is the status of the claim that the mechanism of action in the population is responsible for effectiveness in the individual? Consider the following questions; can both be answered in the affirmative?

**Exemplification.** Are the crucial features of the mechanism of action in the population preserved in the individual?

**Masking.** Are there further mechanisms operating in the individual that counteract the mechanism operating in the population?

Note that if exemplification has been established and masking ruled out, it is possible to particularise a population-level causal claim to an individual without the need for the population to be homogeneous with respect to the mechanism of action. On the other hand, a high degree of homogeneity provides *prima facie* evidence for exemplification and against masking, and thereby supports particularisation.

**Example. Lactose intolerance**

The world population is not very homogeneous with the reaction to milk intake. About 65% of people are lactose intolerant at some point in their lives. However, in different populations there are differing frequencies of lactose intolerant members. Only 5% of Northern Europeans and more than 90% in some populations in East Asia are lactose intolerant, for instance (NIH, 2017). This is because in East Asia lactase deficiency is quite common, while it is quite unusual in Northern Europe. Now, establishing that the patient has no lactase deficiency may be sufficient to establish that she may safely drink milk at high doses. However, even if ruling out lactase deficiency is not possible, establishing homogeneity in a relevant subpopulation may provide grounds for provisionally establishing causality in its members. If, for instance, a patient is North European, this may make it quite plausible that she can drink...
milk safely. If, on the other hand, a patient is East Asian, this may make it quite plausible that she cannot drink milk safely.

**Example. The Shonubi case**

Nigerian drug-mule Shonubi was caught on his eighth trip from Nigeria on the JFK airport carrying heroin in his digestive tract (Colyvan et al., 2001). For sentencing purposes, it was assessed whether the total amount of drugs smuggled on his seven prior trips was greater than a specific amount $M$. There was statistical data available for the amount of drugs carried by balloon-swallowing heroin smugglers from Nigeria. Moreover, there is a social mechanism involving these smugglers that helps to explain the amount of drugs they smuggle: the local drug organisation trains the mules in balloon-swallowing for several weeks and threatens people who refuse with violence (Izenman, 2000).

It seems best to estimate the amount of drugs smuggled by Shonubi on his seven prior trips by the average amount smuggled by balloon-swallowing heroin smugglers from Nigeria. There is high quality mechanistic evidence for application to Shonubi available. Firstly, the mechanism that connects balloon-swallowing heroin smugglers from Nigeria to the quantity of drugs smuggled does apply to Shonubi. The local organisation did indeed train Shonubi by similar methods to those applied to other drug mules, for instance. Secondly, it seems that, for all we know, there is no counteracting mechanism that makes Shonubi an exceptional drug mule. Note that the trip on which he was caught was already his eighth. Thirdly, although there is some variability with respect to the amount smuggled within balloon-swallowing heroin smugglers from Nigeria, virtually all drug mules smuggled more than $M$ grams. Hence, the balloon-swallowing heroin smugglers from Nigeria is arguably a sufficiently homogeneous population.

To obtain the status of effectiveness for a particular individual, one can combine the status of the effectiveness claim in the population with the status of the mechanistic similarity claim (i.e., the claim that there is exemplification and no masking), as in Table 1.15.

A few remarks shed some light on this table.

First, observe that effectiveness in an individual can never be ruled out by the fact that the mechanism responsible for effectiveness in the population is not present in the individual. After all, the individual may exemplify an alternative mechanism of action. I.e., the individual may be a member of a different population, which also exhibits effectiveness but with a different mechanism of action, and this alternative mechanism is present in the individual.
Second, particularisation is a special case of extrapolation. When particularised, a causal claim is extrapolated to the subpopulation of population-members that share all the relevant properties of the individual. This target subpopulation will typically be small, but it remains a subpopulation. Suppose, for instance, we are interested in whether Ralph, a 30 year old Norwegian farmer, will develop an adverse reaction when drinking milk. 95 percent of individuals in Northern Europe show no such reaction. Here, the target population relevant to particularisation may contain only the farmer in question, while the source population is the class of all Northern Europeans.

Third, there are nevertheless some differences between the evaluation of external validity and the evaluation of particularisation to an individual. Particularisation to the individual is more likely to succeed than is extrapolation from a source population to a target population that is not a subpopulation of the source population. This is because causality established in a population is more informative about individuals in this population than about individuals in different populations. For instance, if the population is very homogeneous, then particularisation to the individual is likely to succeed while extrapolation to other populations may well fail. This fact is reflected in the above tables. Consider the case where no studies are available which involve the particular individual. If mechanistic similarity is provisionally established and effectiveness is established in the population, the causal claim is provisionally established for the individual, according to the particularisation table. In the case of external validity, if mechanistic similarity between the source and target populations is provisionally established and effectiveness is established in the source population, effectiveness in the target population is only arguable (see §1.5.2).

Note finally that, in contrast to the method of evaluating external validity in §1.5.2, in the present appendix we treat the case where there is no evidence for causation obtained by studies directly on the target population.

<table>
<thead>
<tr>
<th>Causation in population</th>
<th>Similarity of mechanism in individual and population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
<td>Established</td>
</tr>
<tr>
<td>Provisionally established</td>
<td>Provisionally established</td>
</tr>
<tr>
<td>Arguable</td>
<td>Arguable</td>
</tr>
<tr>
<td>Speculative</td>
<td>Speculative</td>
</tr>
</tbody>
</table>
In this appendix we present a simple probabilistic interpretation of the quality levels and status levels introduced in §1.2.2. This is intended to show that the approach developed in this book is compatible with, rather a rival to, a probabilistic analysis.

Let:

- $E$ be current evidence
- $E'$ be a hypothetical substantially increased future evidence base
- $C$ be the claim under consideration
- $x = P(C|E)$
- $x' = \sum_{E'} P(E')P(C|E')$
- $R(y)$ be a small region around $y$, for any $y$ in the unit interval.

Then the quality levels can be interpreted as in Table 1.16. The status of the claim can be interpreted as in Table 1.17.

### Table 1.16 A probabilistic interpretation of the quality level of evidence.

<table>
<thead>
<tr>
<th>Quality</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Further research is highly unlikely to have a significant impact on our confidence in the claim: $P(x' \in R(x)) \in R(1)$</td>
</tr>
<tr>
<td>Moderate</td>
<td>Further research is moderately unlikely to have a significant impact on our confidence in the claim: $R(1) &gt; P(x' \in R(x)) &gt; 1/2$</td>
</tr>
<tr>
<td>Low</td>
<td>Further research is moderately likely to have a significant impact on our confidence in the claim: $1/2 \geq P(x' \in R(x)) &gt; R(0)$</td>
</tr>
<tr>
<td>Very low</td>
<td>Further research is highly likely to have a significant impact on our confidence in the claim: $P(x' \in R(x)) \in R(0)$</td>
</tr>
</tbody>
</table>

Notes:
1. These levels depend on appropriately ‘small’ intervals $R(y)$. These intervals are intentionally left vague, because the quality levels and status levels are themselves vague categories. The intervals may need to differ from category to category, and differ between the explication of quality and the explication of status.
2. Under a probabilistic interpretation, the tables presented here for determining status, such as the table for external validity in §1.5.2, should be viewed as heuristic approximations to determining status by means of a full probabilistic analysis.
3. A technical point: it is debatable as to whether subjective Bayesianism can make sense of these distinctions. Subjective Bayesianism tends to hold that current probability equals the expectation of future probability, in which case it is hard
Table 1.17 A probabilistic interpretation of the status of a claim.

<table>
<thead>
<tr>
<th>Status</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Established</td>
<td>$x \in \mathcal{R}(1)$ and evidence is high quality, $P(x' \in \mathcal{R}(x)) \in \mathcal{R}(1)$</td>
</tr>
<tr>
<td>Provisionally established</td>
<td>$x \in \mathcal{R}(1)$ and evidence is moderate quality, $\mathcal{R}(1) &gt; P(x' \in \mathcal{R}(x)) &gt; 1/2$</td>
</tr>
<tr>
<td>Arguable</td>
<td>$\mathcal{R}(1) &gt; x &gt; 1/2$ and evidence is at least moderate quality, $P(x' \in \mathcal{R}(x)) &gt; 1/2$, or $x \in \mathcal{R}(1)$ and evidence is low quality, $1/2 &gt; P(x' \in \mathcal{R}(x)) &gt; \mathcal{R}(0)$</td>
</tr>
<tr>
<td>Speculative</td>
<td>A claim is speculative if it falls into none of the other categories</td>
</tr>
<tr>
<td>Arguably false</td>
<td>$\mathcal{R}(1/2) &gt; x &gt; \mathcal{R}(0)$ and evidence is at least moderate quality, $P(x' \in \mathcal{R}(x)) &gt; 1/2$, or $x \in \mathcal{R}(0)$ and evidence is low quality, $1/2 &gt; P(x' \in \mathcal{R}(x)) &gt; \mathcal{R}(0)$</td>
</tr>
<tr>
<td>Provisionally ruled out</td>
<td>$x \in \mathcal{R}(0)$ and evidence is moderate quality, $\mathcal{R}(1) &gt; P(x' \in \mathcal{R}(x)) &gt; 1/2$</td>
</tr>
<tr>
<td>Established</td>
<td>$x \in \mathcal{R}(0)$ and evidence is high quality, $P(x' \in \mathcal{R}(x)) \in \mathcal{R}(1)$</td>
</tr>
</tbody>
</table>

to maintain both a high value of $x$ and a low value of $P(x' \in \mathcal{R}(x))$. This limitation does not apply to objective Bayesianism as developed, e.g., by Williamson (2010).

1.H Glossary

There is a correlation between two variables $A$ and $B$ if these two variables are probabilistically dependent, i.e., $P(B|A) \neq P(B)$. In many situations where a causal relationship is being assessed, the correlation claim of interest is the probabilistic dependence of $A$ and $B$ conditional on some set of a priori potential confounding variables. A confounding variable is a variable correlated with both $A$ and $B$, such as a common cause of $A$ and $B$. Note that ‘correlation’ is sometimes used to refer to a linear dependence; here we use the term in the more general sense to refer to any probabilistic dependence.

A claim of effectiveness is a claim that a particular causal relationship holds in some target population of interest.

A claim of efficacy (or internal validity) is a claim that a particular causal relationship holds in some specific study population.
A claim of **external validity** (or **applicability**) is a claim that a particular causal relationship holds in a population other than, or wider than, a specific study population.

A **clinical study** for the claim that \(A\) is a cause of \(B\) repeatedly measures the values of a set of measured variables that includes the variables \(A\) and \(B\). These values are recorded in a **dataset**. In an **experimental study** the measurements are made after an experimental intervention. If no intervention is performed the study is an **observational study**: a **cohort study** follows a group of people over time; a **case control study** divides the study population into those who have a disease and those who do not and surveys each cohort; a **case series** is a study that tracks patients who received a similar treatment or exposure. An **n-of-1 study** consists of repeated measurements of a single individual; other studies measure several individuals. Clinical studies are crucial for estimating any correlation between \(A\) and \(B\), and they indirectly provide evidence relevant to the claim that \(A\) is a cause of \(B\) (see Figure 1.1).

A **mechanistic study** for the claim that \(A\) is a cause of \(B\) is a study which provides evidence of features of the mechanism by which \(A\) is hypothesised to cause \(B\). Table ?? contains some examples of mechanistic studies. In particular, consider a clinical study for the claim that \(A\) is a cause of \(C\), where \(C\) is an intermediate variable on the mechanism from \(A\) to \(B\)—e.g., a surrogate outcome. Such a study is also a mechanistic study because it provides evidence of certain details of the mechanism from \(A\) to \(B\). A clinical study for the claim that \(A\) is a cause of \(B\) is not normally a mechanistic study for the claim that \(A\) is a cause of \(B\) because, although it can provide indirect evidence that there exists some mechanism linking \(A\) and \(B\), it does not normally provide evidence of the structure or features of that mechanism. Similarly, a mechanistic study for the claim that \(A\) is a cause of \(B\) is not normally a clinical study for the claim that \(A\) is a cause of \(B\), because it does not measure values of \(A\) and \(B\) together. A study will be called a **mixed study** if it is both a clinical study and a mechanistic study—i.e., if it both measures values of \(A\) and \(B\) together and provides evidence of features of the mechanism linking \(A\) and \(B\). To avoid confusion, the terminology **clinical study** and **mechanistic study** will be used to refer only to non-mixed studies.

**General mechanistic claim.** In the case of efficacy, the general mechanistic claim takes the form: there exists a mechanism linking the putative cause \(A\) to the putative effect \(B\), which explains instances of \(B\) in terms of instances of \(A\) and which can account for the observed correlation between \(A\) and \(B\). In the case of external validity, the general mechanistic claim is: the mechanism responsible for \(B\) in the target populations is sufficiently similar to that responsible for \(B\) in the source population.

**Specific mechanism hypothesis.** This is a hypothesis of the form: mechanism \(X\) with features \(F\) links the putative cause to the putative effect.

**Masking mechanism.** A mechanism that counteracts the action of another mechanism.
Quality level of an evidence base. Quality levels are *high, moderate, low* and *very low*. See §1.2.2.

Status of a claim. A claim can be *established, provisionally established, arguably true, speculative, arguably false, provisionally ruled out, or ruled out*. See §1.2.2.

References


Russo, F. and Williamson, J. (2012). EnviroGenomarkers: the interplay between mechanisms and difference making in establishing causal claims. *Medicine Stud-
1  Evaluating evidence of mechanisms in medicine


