Evaluating Evidence of Mechanisms in Medicine: a handbook for practitioners

Veli-Pekka Parkkinen, Christian Wallmann, Michael Wilde, Jon Williamson, Brendan Clarke, Phyllis Illari, Mike Kelly, and Federica Russo

Version 0.1

1 Introduction

Much of medical practice depends upon establishing *effectiveness*.[1] This includes, for example, identifying the causes of cancers, 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.

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. First, that this correlation is genuinely causal in the study population - this is called establishing *efficacy.* Second, that this causal relationship applies outside the study population to the population of interest - i.e., one needs to establish its *external validity*.

As we shall discuss below, evidence of mechanisms is crucial to both of these steps.[2] 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'.

1.1 What is a mechanism?

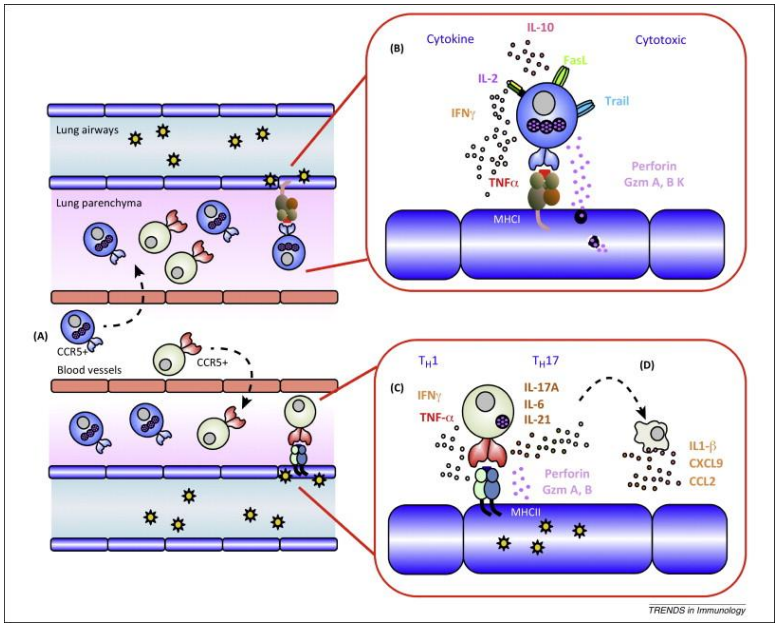
Mechanisms are invoked to explain (Machamer et al. 2000). Textbooks in the biomedical and social sciences are littered with diagrams and descriptions of mechanisms. These are used

[1] With the term 'medical practice' we shall understand, broadly, clinical, scientific, and political forms of engagement with health and disease (Clarke & 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] The view that mechanisms should be considered alongside correlations when establishing causality is implicit in the causal criteria put forward by Bradford Hill (1965). Several of Bradford Hill's indicators of causality are good indicators of mechanisms, while several are good indicators of correlation. The approach of this manual is very much in this tradition. Bradford Hill's approach is widely endorsed today. However, while clinical studies are explicitly scrutinised and evaluated, other evidence of mechanisms is often treated implicitly and intuitively. Our approach is to make evidence of mechanisms explicit, so that it too can be scrutinised and challenged.

to explain the proper function of features of the human body, to explain diseases and their spread, and to explain the functioning of medical devices, 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 & Williamson 2012). In such mechanisms, spatio-temporal and hierarchical organisation tend to play a crucial explanatory role. This can be seen, for example, in a diagram of T cell effector mechanisms in a lung infected by influenza A virus (Gruta & Turner 2014):



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.

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 & 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 can be social as well as biological (Kelly et al 2014), and, in the case of medical devices for instance, they may be neither social nor biological.

1.2 Where does evidence of a mechanism come from?

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. For example (Clarke et al. 2014):

> ***Sources of evidence of mechanisms***
>
> Direct manipulation: e.g., *in vitro* experiments
> Direct observation: e.g., biomedical imaging, autopsy, case reports
> Clinical studies: e.g., RCTs
> Confirmed theory: e.g., biochemistry
> Analogy: e.g., animal experiments
> Simulation: e.g., agent-based models

1.3 Why consider evidence of mechanisms?

Evidence of mechanisms can inform a variety of tasks. In this manual, we focus on its use for evaluating efficacy and external validity.

1.3.1 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 *A* causing *B*, rather than some other explanation (Williamson 2017):

> ***Possible explanations of an observed correlation between A and B:***
>
> **Causation:** *A* is a cause of *B*.
> **Reverse causation**: *B* is a cause of *A*.
> **Confounding (selection bias)**: There is some common cause *C* of *A* and *B* that has not been adequately controlled for by the study.
> **Performance bias**: Those in the *A* group are identified and treated differently to those in the non-*A* group.
> **Detection bias**: B is measured differently in the *A* group in comparison to the non-*A* group.
> **Chance:** Sheer coincidence, attributable to, e.g., too small a sample.
> **Fishing:** Measuring so many outcomes that there is likely to be a (chance) correlation between *A* and some such *B*.
> **Temporal trends**: *A* and *B* both increase over time for independent reasons. E.g., prevalence of coeliac disease - spread of HIV.
> **Semantic relationships**: Overlapping meaning. E.g., phthiasis, consumption and scrofula are correlated because they all refer to TB.
> **Constitutive relationships**: One variable is a part or component of the other.

Logical relationships: Measurable variables A and B are logically complex and logically overlapping, e.g., *A* is *C&D* and *B* is *DvE*.
Physical laws: E.g., conservation of total energy can induce a correlation between two energy measurements.
Mathematical relationships: E.g., mean and variance variables from the same distribution will often be correlated.

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.

More generally, evidence of mechanisms is very useful for determining what explains an observed correlation. It can help rule in or out many of the possible explanations listed in the above box. 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 the above box can be rendered less credible by choosing a particular study design. Matching 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 disjoint can reduce the probability of some of the other explanations.

In certain cases, clinical studies alone can establish that an observed correlation is causal. Possible examples include interventions such as administering aspirin to relieve headache or even general anesthesia, the mechanism of which is not fully understood (see Howick 2011). However, establishing a causal claim in the absence of evidence of mechanisms requires several independent trials 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 trials 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 leads to more reliable causal conclusions. Where this evidence needs to be considered, its quality should be evaluated in ways analogous to the assessment of quality of evidence produced by clinical studies.

1.3.2 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:

- a sufficiently similar mechanism exists in the target population, and
- any further mechanisms there 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.

Note that this form of inference requires even more in the way of evidence of mechanisms than does establishing efficacy (Williamson 2017). When establishing efficacy it is enough to establish the *existence* of a mechanism. On the other hand, when establishing external validity one often needs to identify the mechanism in question (i.e., its entities, activities, processes and organisation) in order to establish that the same mechanism operates in the population of interest, and one also needs to establish something about counteracting mechanisms. 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 a component of an intervention mechanism and may in fact hinder the effectiveness of the intervention.[3]

1.3.3 Other questions

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

- Drawing inferences about a single individual (for treatment and personalised medicine),
- Commissioning new research and devising new research funding proposals,
- Designing clinical trials 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

[3] Well known examples involve the Tamil Nadu Integrated Project (India) and the North Karelia Project (Finland), both discussed by Clarke et al. 2014. The Tamil Nadu Integrated 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.

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: 1505).

This changed with the discovery that the hypersensitivity syndrome only occurred in individuals with the HLA-B*5701 allele (Mallal, Nolan, Witt 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, Nolan, Witt et al 2002: 727). Patients are now genetically screened for the HLA-B*5701 allele, and this has greatly reduced the incidence of the hypersensitivity syndrome (Rauch, Nolan, Martin et al 2006).

In the present version of this document, we focus on the use of evidence of mechanisms to help establish efficacy and external validity. Some of these other questions will be addressed in future versions of this document.

2 How to consider evidence of mechanisms: a summary

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

2.1 Questions to address

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

Efficacy

Does the effect size and quality of clinical studies establish that the observed correlation is causal?

No?

- Evaluate non-clinical-study 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 non-clinical-study 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?

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. Moreover, regarding 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. 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.

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.

2.2 Quality levels

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). In each case, the ranking will be measured on the following scale:

Quality level	**Interpretation**
High	Further research is highly unlikely to have a significant impact on our confidence in the truth of the claim. If this level of confidence is sufficiently high - i.e., if high-quality evidence renders the claim sufficiently plausible - the claim is **established**, in the sense that community standards are met for adding the claim to our body of evidence. If the level of confidence is sufficiently low - i.e., if high-quality evidence renders the *falsity* of the claim sufficiently plausible - the claim is **ruled out**, in the sense that its negation can be added to our body of evidence.
Moderate	Further research is moderately unlikely to have a significant impact on our confidence in the truth of the claim. If the level of confidence is sufficiently high, the claim is **provisionally established** or **provisional**. If the level of confidence is sufficiently low, the claim is **provisionally ruled out**.
Low	Further research is moderately likely to have a significant impact on our confidence in the truth of the claim. If the claim is clearly more plausible than not then it is **arguably true** or **arguable**. If the claim is clearly more plausibly false than true, it is **arguably false**.
Very low	Further research is highly likely to have a significant impact on our confidence in the truth of the claim. The claim is **speculative**.

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.) Finally, note that the above table invokes two separate levels: the quality level applies to the evidence, while the level of confidence applies to the claim in question. As to whether a claim is established depends on both the quality of the evidence as well as the degree of confidence that the evidence warrants. We shall refer to the category *established, provisionally established*, …, *ruled out* as the status of the claim in question.

This table of quality levels is very close to the original GRADE approach to assessing quality of evidence for a correlation, put forward by Guyatt et al. (2008).[4] 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).

2.3 Identifying evidence of mechanisms in the literature

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, there are a large number of such studies, and the database indexing practices for these studies tend to be unsystematic (Smith et al 2016). The studies tend to lack standardized search terms in their titles and abstracts. In general, the titles, abstracts, and indexing of clinical studies have become standardized in a way that facilitates literature searches for such studies (Evans 2002). For the most part, this has not yet happened for mechanistic studies. This has led to a tendency to overlook or entirely ignore evidence of mechanisms that arises from sources other than clinical studies.

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

1. Scoping and identifying a general mechanism hypothesis;
2. Formulating a number of review questions on the basis of this hypothesis, and using these review questions to search the literature;

[4] 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). 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 *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. Plausibly, confidence should remain at 75%. Arguably, however, this confidence is not sufficiently stable for the claim to be considered established or even provisionally established.

3. Identifying the evidence most relevant to the mechanism hypothesis by refining the results of this search;
4. Presenting the evidence of mechanisms.

This strategy, developed in detail in Section 3, helps to overcome some of the practical difficulties with identifying evidence of mechanisms, difficulties which may prevent practitioners from considering the evidence of mechanisms required for establishing efficacy and external validity.

2.4 Evaluating evidence of mechanisms

In evaluating the quality of an item of mechanistic evidence, one should consider the following questions.

> 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 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.
>
> Can the item of evidence be produced by many independent methods? Employing many 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.)
>
> Are the model systems that are used in experimental research representative of humans? The more faithfully the model systems reproduce the relevant human features, the higher the quality of evidence gleaned from them.
>
> Can the mechanism be observed operating in many different background contexts? The more robust a mechanism, the more reliable the inferences we make based on the mechanistic evidence. Demonstrable robustness of the mechanism itself thus makes for higher quality evidence.

The status of a mechanistic claim (c.f. Section 2.2 Quality levels) can be assigned as follows. A mechanism to account for efficacy is considered established when either high quality trial evidence - not explainable by, e.g., confounding or bias - exhibits a substantial correlation, or when high quality analytic evidence confirming all the crucial component features of the mechanism is available. 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 controlled trials 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 for the similarity of all the crucial components of the mechanism in the study and target populations is available. 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. Arguably true claims have evidence in their support that is either low quality or gappy in some important respects. Speculative claims are supported by evidence that show mixed results, or have little evidence in their support beyond theoretical intuition or speculation.

These questions are explained in more detail in Section 4.

2.5 Using evidence of mechanisms to evaluate causal claims

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 Section 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 correlational claim and the status of the mechanistic claim. For instance, if a correlation is arguable but a 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 study population, and (iii) evidence of similarity of mechanisms between study and target populations. Evidence directly about the target may be boosted (respectively, undermined) by observing that efficacy does (respectively, does not) hold in a study population that shares similar mechanisms with the target population. The following table 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:

SOURCE: TARGET: Causation in target pop. on the basis of evidence directly about the target:	Established causation in source + established similarity of mechanism	Provisionally established + established	*Other combinations*	Provisionally ruled out causality in source + established similarity, or ruled out causality in source + provisionally established similarity	Ruled out causality in source + established similarity of mechanism
Established	Established	Established	Established	Established	Provisionally established
Provisionally established	Established	Provisionally established	Provisionally established	Provisionally established	Arguable
Arguable	Established	Provisionally established	Arguable	Speculative	Speculative
Speculative	Established	Arguable	Speculative	Arguably false	Ruled out
Arguably false	Speculative	Speculative	Arguably false	Provisionally ruled out	Ruled out
Provisionally ruled out	Arguably false	Provisionally ruled out	Provisionally ruled out	Provisionally ruled out	Ruled out
Ruled out	Provisionally ruled out	Ruled out	Ruled out	Ruled out	Ruled out

Having summarised the overall approach, we now turn to a more detailed consideration of the last three subsections.

3 Identifying evidence of mechanisms in the literature

This section suggests a four-step strategy for identifying evidence of mechanisms.

1. Propose a general mechanism hypothesis as a result of a scoping stage.
2. Formulate a number of review questions on the basis of this hypothesis, and use these review questions to search the literature.
3. Identify the evidence most relevant to the mechanism hypothesis by refining the results of this search.
4. Present this evidence of mechanisms in a clear manner.

Existing evidence of mechanisms can help at each step of this strategy. This strategy helps to overcome some of the practical difficulties with identifying further evidence of mechanisms, difficulties which may prevent practitioners from considering the necessary evidence of mechanisms.

3.1 The scope

The scope provides general information about the intervention under consideration, the relevant population, comparators, health outcomes, a brief overview of the available evidence, and the key problems that will need to be considered in determining effectiveness or external validity. At this stage, it is useful to consider the following question: *Is explicitly considering evidence of mechanisms likely to help in determining the effectiveness of the intervention in the target population?*

It may be that the scope makes it clear that it is unnecessary to explicitly identify evidence of mechanisms in the literature. In this way, existing evidence of mechanisms may help the scoping stage by focusing in on the more relevant issues.

- **Example:** The overview of the evidence may make clear that the trial evidence alone will be sufficient to establish the effectiveness of the intervention.
- **Example:** The overview may make clear that background knowledge of relevant mechanisms is enough to establish effectiveness.

In other cases, however, it may be that the key problems identified at this scoping stage are addressed by explicitly considering evidence of mechanisms.

- **Example:** The overview of the evidence may make it clear only that there is likely a correlation between the intervention and health outcome of interest in the population. In this case, a key problem will be to establish whether this correlation is causal.
- **Example:** The overview of the evidence may suggest that an intervention causes a health outcome only in a study population. In this case, a key problem will be establishing the extent to which there are similar mechanisms in the study and target populations.

If evidence of mechanisms is likely to be helpful, the next step is to *identify a general mechanism hypothesis*.

- **Efficacy:** In the case of efficacy, there is a single crucial question: *What mechanism may be proposed to account for an observed correlation between an intervention and a health outcome in the study population?* A general mechanism hypothesis is a hypothesis put forward to answer a question such as this.
- **External validity:** In the case of external validity, it may seem straightforward to identify the relevant mechanism hypothesis: *Is the mechanism that accounts for an observed correlation between an intervention and a health outcome in the study population also present in the target population?* However, it is important to consider also the possibility of further mechanisms in the target population that may affect the extent of the correlation between the intervention and health outcome: *Are there any masking mechanisms in the target population?*

There are a number of ways to identify a general mechanism hypothesis:

- A mechanism hypothesis may be proposed on the basis of the **clinical study literature**.
  - If a clinical study establishes a correlation between an intervention and a health outcome, and the suggestion is that this correlation is causal, then the authors of such a study usually at least propose a general mechanism hypothesis of the following form: *There is a mechanism X linking the intervention and the health outcome in the study population*. Some of these mechanisms may also be masking mechanisms.
  - Given this, the discussion section of a paper reporting the results of a clinical trial is a good place to look in order to identify a general mechanism hypothesis.

Example: The discussion section of a recent paper on the effect of long-term aspirin use on the risk of cancer says: '[O]ur 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 proinflammatory prostaglandins, including prostaglandin E2, which increases cellular proliferation, promotes angiogenesis, and increases resistance to apoptosis. Aspirin may also play a role in Wnt signaling, nuclear factor κB signaling, polyamine metabolism, and DNA repair' (Cao et al 2016). References are given for these hypothesized mechanisms.

- A mechanism hypothesis may also be proposed on the basis of the **basic science literature** or **clinical expertise**.
  - The existing evidence from mechanistic studies or clinical expertise may be sufficient to hypothesize a mechanism.

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 general mechanism linking a reduction in the size of obstructions in the airway 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, viz., 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).

When evidence is to be evaluated by a committee of experts (as often happens at NICE, IARC, MHRA, CHMP, for instance), it is useful to provide a list of plausible mechanisms 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.

3.2 Formulate the review questions, and review the literature.

The mechanism hypothesis will have implications that concern the entities, activities, and their organization in the mechanism. These implications may be used to formulate a number of specific review questions with which to review the literature for evidence relevant to the mechanism hypothesis. The mechanism hypothesis receives some confirmation if these implications are observed, and it is disconfirmed if observations conflict with the implications of the hypothesis. In addition, the review provides information about the details of the mechanism which is useful information for determining external validity.

- Determine the **implications** of the general mechanism hypothesis for the putative entities, activities, and their organization.
    - Existing evidence of mechanisms may be useful in determining these implications.

Example: *The ten key characteristics of carcinogenicity.* 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. 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.

A correlation between some environmental exposure and a type of cancer in a study population may be observed, and this may lead to the proposal of the general mechanism hypothesis, viz., that there is a mechanism linking the environmental exposure and the cancer. But this general mechanism hypothesis can lead to more specific review questions involving implications of the hypothesis about entities, activities, and their organization, by relying on existing evidence of the mechanisms by which exposures frequently lead to cancer. See Smith et al (2016).

- Use these implications to inform a number of specific **review questions** with which to search the literature.
    - The review questions will inform the evidence review, and so should be clear and precise. The evidence review will determine the most suitable sources of evidence. But it may also be revised in the light of new evidence.
    - 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 review question. It is those crucial and contentious features of the proposed mechanism that should figure in the review question.

- Identify research potentially relevant to the assessment of the hypothesized mechanisms by looking at the relevant **non-clinical-study literature**.
    - In the first instance, this may be done by following up the references from the discussion section of the clinical trial report which proposes the causal mechanism as the best explanation of a correlation. Any other publicly available reports may be useful here also, e.g., government agency reports, doctoral theses, etc.
    - More systematically, a preferred method for searching the literature may be used. E.g., a PubMed search---http://www.ncbi.nlm.nih.gov/pubmed---using appropriate MeSH terms, including key terms from the hypothesized mechanisms.
    - See, for example, also the following table of relevant databases:

ConsensusPathDB-human http://consensuspathdb.org/	Database of protein-protein, genetic, metabolic, signaling, gene regulatory and drug target interactions
DisGeNET http://www.disgenet.org/	Database of gene-disease associations
GeneGo MetaCore (curated) https://portal.genego.com/	Search tool for pathway analysis based on curated database of omics data, also gene-disease associations
GWAS catalog http://www.ebi.ac.uk/gwas/	Database of SNP-trait associations

The Binding Database https://www.bindingdb.org/	Search engine for empirically verified binding affinities of drug targets, mostly proteins
KEGG pathway database http://www.genome.jp/kegg/pathway.html	Molecular interactions, reactions and relations
MACiE - Mechanism, Annotation and Classification in Enzymes MACIE	Database of enzyme reaction mechanisms
http://metacyc.org/	Metabolic pathway database
Online Mendelian Inheritance in Man OMIM http://www.omim.org/	Database of human genetic disorders
PubChem	Search tool for structure and bioactivity of small molecules
Reactome.org (curated)	Curated, open source searchable database of human physiological pathways of any type
http://string-db.org/	Database of known and predicted protein-protein interactions
SuperCYP	Database of >3000 drug responses metabolized by Cytochrome P450 enzymes

3.3 Identify the evidence most relevant to the mechanism hypothesis.

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

- Use preferred inclusion and exclusion criteria and expert knowledge to rule out evidence that is not relevant or that is of insufficient quality (Kushman et al 2013).
    - For example: *Does the publication include original data*? A good rule of thumb: if it does not include original data, then exclude the publication.
- There are content management tools available to help in identifying, screening, organizing, and summarizing the evidence.
    - For example: Health Assessment Workspace Collaborative (HAWC), https://hawcproject.org/.

Example: It seems to be established that aspirin works to modify COX enzymes. The vascular benefits of aspirin understood in terms of COX enzyme generated products seems 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'. But one could make up one's own mind by following up their references, or searching the literature with their terms.

3.4 Presenting the evidence of mechanisms

It is important to present a clear summary of the identified evidence of mechanisms. This will make it more straightforward to consider the quality of that evidence, that is, the extent to which the overall evidence of mechanisms makes it likely that there exists a mechanism underlying a given correlation. (Presenting the quality of evidence of mechanisms is a separate issue, for which guidance is provided in section 4.2.) A summary of evidence of mechanisms should clearly state what outcome the mechanism in question is proposed to account for, that is, whether it is presented as evidence of a mechanism of action to account for the efficacy of a treatment, or as evidence of similarity of mechanisms between populations to account for external validity of an effect demonstrated in a study population. For presentation purposes, the evidence can be summarized as a theoretical, narrative, explanation for the claim of interest. Often, such an explanation can be conveniently summarized in a diagram as well. Such a presentation should clearly indicate key features of the mechanism: the properties of the individual components, and their organization, hypothesized to be responsible for the outcome of interest.

4 Evaluating evidence of mechanisms

Introductory remarks. Evidence of mechanisms is needed to establish that an observed correlation is causal, and to evaluate the properties of an established causal relation, such as its sensitivity to changes in background conditions or parts of the mediating mechanism. The former use corresponds to establishing efficacy in a study population, while the latter is relevant for evaluating the external validity of results. For establishing efficacy, one needs to know that a mechanism mediating the putative cause and effect exists - it is not necessary to know the features of the mechanism in full detail. For establishing external validity, one needs to establish that the mechanisms in study and the target populations are sufficiently similar. This inference might proceed in one of the following ways.

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 study population, and then inferring that the mechanisms in the study and target population are similar due to shared ancestry of the populations. The greater the degree of isolation between the target population and the population from which the study population was sampled, the less reliable this inference will be.

For establishing efficacy, one can establish that a mechanism exists by observing its overall effect in a controlled experiment, without necessarily knowing the details of how the mechanism produces the effect. Robust evidence of a sufficiently large correlation between the independent and dependent variables gleaned from well conducted randomized controlled trials can be enough to establish that some mechanism linking the variables exists. Good trial evidence can thus be high quality evidence for the *existence* of a mechanism. Another strategy is to proceed analytically, by evaluating the evidence for each component feature of the mechanism - each process, entity, activity and organisational feature - and then putting the evidence of these component features together in order to evaluate the evidential support for the mechanism as a whole. If one's analytic understanding of a mechanism involves many "black boxes" - merely hypothesized components or organizational features - the evidence for the mechanism overall should be judged as low quality. If all or most of the component features and their organization can be supported with strong evidence, and the mechanism can be experimentally shown to produce the phenomenon of interest, the evidence of mechanism should be judged high quality. Most cases where evidence of mechanism is at issue will fall between these two extremes.

In practice, if one can consult a textbook or a curated database for a summary of evidence of a well established mechanism, there may be no need to go through the process of evaluating evidence for each component feature of a mechanism individually. However, often some parts of the mechanism are better understood than others, and a piecemeal evaluation of the evidence base is required when a new treatment targeting a novel part of the mechanism is proposed. This section describes the criteria by which evidence of mechanism is evaluated, and presents a system for grading evidence of mechanisms for different

purposes: establishing (or ruling out) mechanisms for efficacy, and establishing (or ruling out) mechanisms that could account for the external validity.

Discovering features of mechanisms typically proceeds by one or more of the following two means:

1. Experimentation: 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, randomization procedure (when applicable) and statistical inference can thus be applied to evaluate the possibility of experimental error. Also, simulation experiments can be used, especially to investigate whether the hypothesized organization 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 organization of a mechanism can be found by observation techniques such as imaging technologies, autopsy, 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).

The particular challenges for evaluating evidence of mechanisms in medicine stem from the fact that evidence of mechanisms 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). In addition, 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, weighing evidence of mechanisms requires evaluating how well these problems have been mitigated in the process of creating the evidence.

4.1 Considerations for evaluating evidence of mechanism

The following considerations are to be used to evaluate the quality of evidence for mechanistic claims. Not all of them are applicable in every case of evidence appraisal, and there is no ranking in terms of importance between them. Instead, evidence that is judged to be of high (respectively, low) quality in the light of several considerations, when applicable, ought to be taken as higher (respectively, lower) quality than evidence ranked of high (respectively, low) quality in the light of just one. For instance, evidence that is robustly reproducible by many well understood methods is to be judged of higher quality than evidence that can be produced by just one well understood method.

Well understood methods and model systems. In order to evaluate evidence of mechanisms 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.

The degree to which experimental systems replicate human features of interest. 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 - such as a cultured tissue, cell population or an animal model - 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 characterized, 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 (of disease or treatment-efficacy) 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.

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

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 to humans.

Based on the amount and quality of evidence, claims about mechanisms - either about the existence of a mechanism in a study population, or about the similarity of mechanisms between study and target populations - are attributed a quality level according to the levels introduced in section 2.2 Quality levels. Note that different types of claim need to be considered for the purpose of evaluating evidence of a mechanism that can account for efficacy, and for the purpose of evaluating a mechanism that can account for external validity. In the former case, one considers the question of whether a mechanism capable of accounting for the correlation exists. 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 (respectively, low) quality level for either type of claim about a mechanism. 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 quality levels. 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 a quality level.

Checklist of questions to consider in evaluating evidence of mechanisms for internal validity

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

- Has a correlation of the same size has been established in many studies under slightly varying circumstances (robust detectability)?

- Is the observed correlation so large that it could not plausibly be explained by bias or confounding, but only by the existence of a mechanism responsible for the correlation?

- Is the mechanism known in some detail? Can it account for the correlation and its size? Is there good quality analytic evidence for most of the crucial features of the mechanism?

- Is it plausible that the behaviour of the mechanism crucially depends on just some component(s) or organizational features? If yes, are such critical features well established? 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?

- Is there at least moderate quality evidence of a plausible counteracting mechanism? If so, does this evidence suggest that the correlation the mechanism is posited to explain is spurious? (If the existence of a mechanism is inferred from correlational evidence, discovering that the 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.

- Is there at least moderate quality evidence that the mechanism exhibits such complexity that its overall behaviour is very unpredictable?

- Is the evidence of mechanisms indirect, i.e., is the hypothesized mechanism inferred from evidence of an analogous mechanism in some other domain?

Checklist of questions to consider in evaluating evidence of mechanisms 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?

- Has a correlation of the same size been established in many studies under slightly varying circumstances, and in many populations related to the target population (robust detectability) 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?

- Is the mechanism known in some detail both in the study population and the target population, and found to be similar in both, and such that it can account for the correlation observed in the study population?

- When the behaviour of the whole mechanism crucially depends on some component(s) or an organizational feature, are the critical features of the mechanism similar in the study and the target populations? If yes, 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?

- Is there at least moderate quality evidence of a plausible counteracting mechanism? Does this evidence suggests 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.

- Is there at least moderate quality evidence of dissimilarity of the mechanisms in the study and the target populations?

- Is there at least moderate quality evidence that the mechanism exhibits such complexity that its overall behaviour is unpredictable?

- Is the evidence of mechanism indirect, i.e., the hypothesized mechanisms are inferred from evidence of an analogous mechanisms in some other domain?

Mechanistic evidence for efficacy or external validity should be evaluated considering the correlation 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 to account in assessing the status of a mechanistic claim. In evaluating a mechanistic claim, evidence arising from clinical studies

and evidence arising from other sources have mutually supporting roles. (This consideration is explored further in section 5.1 Efficacy.

4.2 Presenting quality of evidence of mechanisms

Preparing and presenting summaries of the quality of mechanistic evidence in a standardized 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 guidelines can nonetheless be given.

The quality of the overall evidence of a mechanism should be presented in such a way that it outlines the quality of the evidence for each of the individual component features of the mechanism, evaluated employing the considerations for evaluating evidence described in section 4.1 Considerations for evaluating evidence of mechanism.

For example, presume that a drug is hypothesized 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 paradigms, that are able to confirm the hypothesized interaction, the higher the quality of evidence for this particular feature of the hypothesized 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. The quality of evidence for particular features of the mechanism can be highlighted in a diagram or discursive summary of the evidence, using symbols to indicate the status of particular features of the proposed mechanism:

Status	Symbol
Established	*
Provisionally established	++
Arguable	+
Speculative	?
Arguably false	-
Provisionally ruled out	--
Ruled out	#

A brief verbal explanation can be included, e.g. ‘++; inconsistencies’.

The quality of the evidence for the overall claim about a mechanism should then be presented, with a brief explanation of how the quality of evidence for the component features of the mechanism justifies the overall quality level. One should be cautious in not overstating the explanatory power of the mechanism beyond what is warranted by the evidence. The

overall quality of the evidence can be summarized in a table or in an annotated diagram, using the same symbols given in the above table.

5 Using evidence of mechanisms to evaluate efficacy and external validity

In this section, we move from mechanism claims to causal claims, i.e., claims of efficacy and external validity. As we have seen, in order to establish efficacy, one normally 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. Section 5.1 shows 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 other populations than the study population (target populations). 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 Section 5.2.

5.1 Efficacy

Here we address the question of how to combine evaluations of a mechanistic claim and a correlation claim in order to evaluate a claim of effectiveness.

Mechanism claim. On the one hand, 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 hypothesized mechanism and is the existence of the crucial features of *this* mechanism established?

Correlation claim. On the other hand, the GRADE system can be used to evaluate whether there is a correlation between the putative cause and effect conditional on all other causes of the effect (Guyatt et al, 2008). It initially awards 4 points to evidence obtained by randomised controlled trials (RCTs) and 2 points to evidence obtained from observational studies. According to standards of quality, consistency, directness and effect size, additional points are awarded or removed (upgrading or downgrading). The quality of the evidence may be considered to be high (respectively, moderate, low, very low) if the total number of points is 4 or more (respectively, 3, 2, 1).

Efficacy claim. To obtain a quality level for efficacy, we combine the quality level of evidence of mechanisms and the quality level of evidence of correlation. According to the Russo-Williamson Thesis (RWT) efficacy is established if and only if it is established that there is a correlation and that there is some mechanism which can account for this correlation (Russo and Williamson, 2007). According to the Generalised Russo-Williamson Thesis, the status of the causal claim is the minimum status of the correlational claim and the mechanistic claim. Hence, a causal claim cannot be better established than one of its necessary conditions. 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.

It can be useful here to 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 evidence of mechanisms and evidence of correlation produces much stronger overall confirmation than would either type of evidence on its own, because they compensate for each other's weaknesses. For instance, having high quality evidence of a correlation of a certain size between putative cause and effect may rule out masking. (Recall that masking occurs when there is one or more counteracting mechanism that cancels out the effect of the mechanism of action.) On the other hand, high quality evidence of mechanisms can rule out confounding.

In the framework set out above, evidence of correlation and evidence of mechanisms reinforce each other in two respects. First, the status of an efficacy claim depends on both the status of a mechanism claim and the status of a correlation claim: a high status for one is not sufficient to confer a high status to the efficacy claim. Second, evidence of correlation may influence the status of mechanistic evidence and vice versa. The following three scenarios consider applications of 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 speculative evidence 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, according to the Generalised Russo-Williamson thesis, efficacy is only speculative. However, this misrepresents the quality levels of the available evidence. It confuses mechanistic evidence obtained *only* by bench research with *total* evidence for mechanisms from all sources. Recall from Section 4.1 Considerations for evaluating evidence of mechanism that clinical studies may also yield evidence relevant to the claim that there exists a mechanism. In the above example, the correlational evidence, when combined with the bench research, can yield a status higher than speculative - an application of the reinforced concrete metaphor. Accordingly, the efficacy claim will have a status higher than speculative.

Scenario 2. Suppose there is high quality evidence that there is a mechanism by which the putative cause can influence the effect. However, the possibility of a further counteracting mechanism cannot be ruled out. In this case, it is not clear that there is a net effect, and evidence of mechanism as a whole may be classified as low quality. Subsequently, high quality correlational evidence is obtained to estimate the relevant net correlation. If this correlation is positive, then this fact provides evidence that any counteracting mechanism fails to totally mask the effect of the mechanism of action. This may now suffice to establish the relevant mechanism claim, namely that there is a mechanism can account for the observed correlation (see Section 4.2 Presenting quality of evidence of mechanisms). Hence, the correlation is taken into account when evaluating the evidence of mechanism - one reinforces the other.

Scenario 3. One might think that low quality evidence of a correlation is subject to confounding, so when there is high quality evidence of mechanism that rules out confounding, efficacy is established. If so, this would contradict the Russo-Williamson Thesis. 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. This accords with the Russo-Williamson Thesis.

Example: There are a number of cases in which a causal claim has been established on the basis of mechanistic studies alone. Arguably, for instance, no clinical studies were necessary to establish the efficacy of the Heimlich maneuver. In these cases, the evidence from the mechanistic studies was sufficient to establish both the existence of a mechanism and the existence of a correlation.

Example: There are cases in which a causal claim is established on the basis of clinical studies alone. Arguably, for instance, deep brain stimulation has been shown to be effective in treating Tourette's syndrome, even though the mechanism by which it helps is not completely understood. In these cases, the clinical studies were sufficient to establish both the existence of a correlation and the existence of a mechanism.

5.2 External validity

If the mechanisms within the study population and the target population are sufficiently similar, this permits the extrapolation of efficacy from the study 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 similar study population.

An initial status of efficacy in the target is given by studies that directly involve the target population. The status can be determined by considering Section 5.1 Efficacy. To obtain a final status for efficacy in the target, we combine this initial status with the status in a study population, provided that study and target population share similar mechanisms. The status of efficacy in the study population can be determined by considering Section 5.1 Efficacy. The status of the mechanistic claim relevant to external validity can be determined as indicated in Section 4.1 Considerations for evaluating evidence of mechanism. The status of the claim about efficacy in the target may be increased (decreased) by observing that efficacy does (not) hold in a study population that is similar to the target population. In this case, claims about efficacy are extrapolated from the study population to the target population.

To change the initial status of an efficacy claim given by studies directly on the target population, further evidence needs to be of high quality. The relevant claim about mechanisms needs to be either established or provisionally established. The evidence of efficacy in the study population needs to be of at least of moderate quality level (c.f., 2.2

Quality levels; the claim needs to be established, provisionally established, provisionally ruled out or ruled out). Other quality levels do not change the initial status.

Table for determining the status of a claim of efficacy in the target population:

SOURCE: TARGET: Causation in target pop. on the basis of evidence directly about the target:	Established causation in source + established similarity of mechanism	Provisionally established + established	*Other combinations*	Provisionally ruled out causality in source + established similarity, or ruled out causality in source + provisionally established similarity	Ruled out causality in source + established similarity of mechanism
Established	Established	Established	Established	Established	Provisionally established
Provisionally established	Established	Provisionally established	Provisionally established	Provisionally established	Arguable
Arguable	Established	Provisionally established	Arguable	Speculative	Speculative
Speculative	Established	Arguable	Speculative	Arguably false	Ruled out
Arguably false	Speculative	Speculative	Arguably false	Provisionally ruled out	Ruled out
Provisionally ruled out	Arguably false	Provisionally ruled out	Provisionally ruled out	Provisionally ruled out	Ruled out
Ruled out	Provisionally ruled out	Ruled out	Ruled out	Ruled out	Ruled out

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 efficacy in the target. If there is a study population for which similarity to the target has been established and efficacy has been ruled out, then causality 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). Appendix A. External validity and IARC relates the above table to the IARC protocol.

2) Changing the status of an efficacy claim obtained from evidence gathered on the target population is more common when that evidence is of lower quality. For instance, the *provisionally established* status may be changed to established 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 mechanism and provisionally established similarity.
3) The GRADE working group also evaluates whether evidence from a study population can be used to draw inferences about the target population. In particular, the GRADE working group considers the case where no evidence directly obtained on the target population is available.

   "In general, one should not rate down for population differences unless one has compelling 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 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, in the human case, the GRADE working group considers 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 for extrapolation too low (see Section 4.1 Considerations for evaluating evidence of mechanism for discussion). In the case of animal studies, the default assumption of the GRADE working group is that efficacy is arguably true if efficacy in animals is established. Again this is problematic. In our approach, in the absence of evidence of similarity of mechanisms, efficacy in the study population cannot be extrapolated to the target. Hence, even if many high quality RCTs in animals establish efficacy in animals, in the absence of further evidence, nothing can be concluded about efficacy in humans.

4) Efficacy can also be established or ruled out even in the case where no evidence in the target is available. This is the case when efficacy has been established in a study population for which it has been established that it is mechanistically similar to the target population.

6 Glossary

A **(clinical) study** or **trial** produces several measurements of the values of a set of measured variables. These 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.

Reference class problem: Suppose that a particular patient belongs to multiple populations for which information about correlations of the dependent variable with the independent variable are available. The reference class problem is the problem of determining the most relevant population for the value of the dependent variable of the individual in question.

Mechanism hypothesis. A hypothesis of a mechanism.

Mechanistic studies are those studies from which evidence of mechanisms is most *characteristically* produced, e.g., bench research, etc. Although evidence of mechanisms can be produced through clinical studies, these *characteristically* produce evidence of correlation.

Masking mechanism: A mechanism that counteracts the action of another known mechanism.

Quality level of an evidence base. Quality levels are *high, moderate, low* and *very low.* See 2.2 Quality levels.

Russo-Williamson Thesis. According to this thesis, establishing efficacy requires establishing both a correlation and the existence of a mechanism that can explain this correlation.

Generalised Russo-Williamson Thesis. According to this thesis the status of an efficacy claim is the minimum of the status of the claim that the variables are appropriately correlated and the status of the claim that there is a mechanism that can account for the correlation and its size. See Section 5.1 Efficacy.

Status of a claim. A claim can be *established, provisionally established, arguably true, speculative, arguably false, provisionally ruled out, or ruled out.* See 2.2 Quality levels.

7 References

Howard Balshem, Mark Helfand, Holger J. Schunemann , Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris, Gordon H. Guyatt 2011. GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology* 64:401-406.

Bonnema, S. J., V. E. Nielsen, H. Boel-Jorgensen, P. Grupe, P. B. Andersen, L. Bastholt, and L. Hegedus. 2008. “Recombinant Human Thyrotropin-Stimulated Radioiodine Therapy of Large Nodular Goiters Facilitates Tracheal Decompression and Improves Inspiration.” Journal of Clinical Endrocrinology and Metabolism 93 (10): 3981–84.

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

Cao et al (2016) Population-wide Impact of Long-term Use of Aspirin and the Risk for Cancer: http://oncology.jamanetwork.com/article.aspx?articleid=2497878

Chan et al (2011) Aspirin in the chemoprevention of colorectal neoplasia: http://cancerpreventionresearch.aacrjournals.org/content/5/2/164.full.pdf+html

Clarke, B.; Gillies, D.; Illari, P.; Russo, F. & Williamson, J. Mechanisms and the Evidence Hierarchy. *Topoi,* **2014***, 33*, 339-360

Clarke B. and Russo F. (2016) Causation in medicine. In *Companion to Contemporary Philosophy of Medicine*. Edited by J. Marcum. Bloomsbury. Ch. 12.

Clay PG . The abacavir hypersensitivity reaction: a review. Clin Ther 2002;24:1502-14.

Evans, D. (2002) Database searches for qualitative research. *Journal of the Medical Library Association*. 90(3). 290-293.

Gruta, N. L. L. & Turner, S. J., T cell mediated immunity to influenza: mechanisms of viral control. *Trends in Immunology,* **2014***, 35*, 396-402

Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, Holger J Schünemann and for the GRADE Working Group 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336;924-926

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., ... & Akl, E. A. (2011). GRADE guidelines: 8. Rating the quality of evidence—indirectness. *Journal of clinical epidemiology*, *64*(12), 1303-1310.

Howick J. (2011) Exposing the vanities-and a qualified defence-of mechanistic reasoning in health care decision-making. *Philosophy of Science* 78(5), 926-940.

Illari, P. M. & Williamson, J., What is a mechanism? Thinking about mechanisms *across* the sciences. *European Journal for Philosophy of Science,* **2012***, 2*, 119-135

Kelly M.P, Kelly R.S., Russo F. (2014) The integration of social, behavioural, and biological mechanisms in models of pathogenesis. Perspectives in Biology and Medicine, 57(3), 308-328.

Kushman ME, Kraft AD, Guyton KZ, Chiu WA, Makris SL, Rusyn I. A systematic approach for identifying and presenting mechanistic evidence in human health assessments. *Regulatory toxicology and pharmacology : RTP*. 2013;67(2):10.1016/j.yrtph.2013.08.005. doi:10.1016/j.yrtph.2013.08.005.

Machamer, P.; Darden, L. & Craver, C., Thinking about mechanisms. *Philosophy of Science,* **2000***, 67*, 1-25

Mallal S, Nolan D, Witt C, et al . Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet 2002;359:727-32.

Nielsen, V. E., S. J. Bonnema, H. Boel-Jorgensen, P. Grupe, and L. Hegedus. 2006. “Stimulation with 0.3-mg Recombinant Human Thyrotropin prior to Iodine 131 Therapy to Improve the Size Reduction of Benign Nontoxic Nodular Goiter: A Prospective Randomized Double-Blind Trial.” Archives of Internal Medicine 166 (14): 1476–82.

Rauch A, Nolan D, Martin A, et al . Prospective genetic screening decreases the incidence of abacavir hypersensitivity reactions in the Western Australian HIV cohort study. Clin Infect Dis 2006;43:99-102.

Russo, F. & Williamson, J, Interpreting causality in the health sciences. *International studies in the philosophy of science* 21.2 (2007): 157-170.

Russo, F. & Williamson, J., EnviroGenomarkers: the interplay between mechanisms and difference making in establishing causal claims. *Medicine Studies: International Journal for the History, Philosophy and Ethics of Medicine & Allied Sciences,* **2012***, 3*, 249-262

Russo, F. & Clarke, B. Mechanisms in medicine. In TBC.

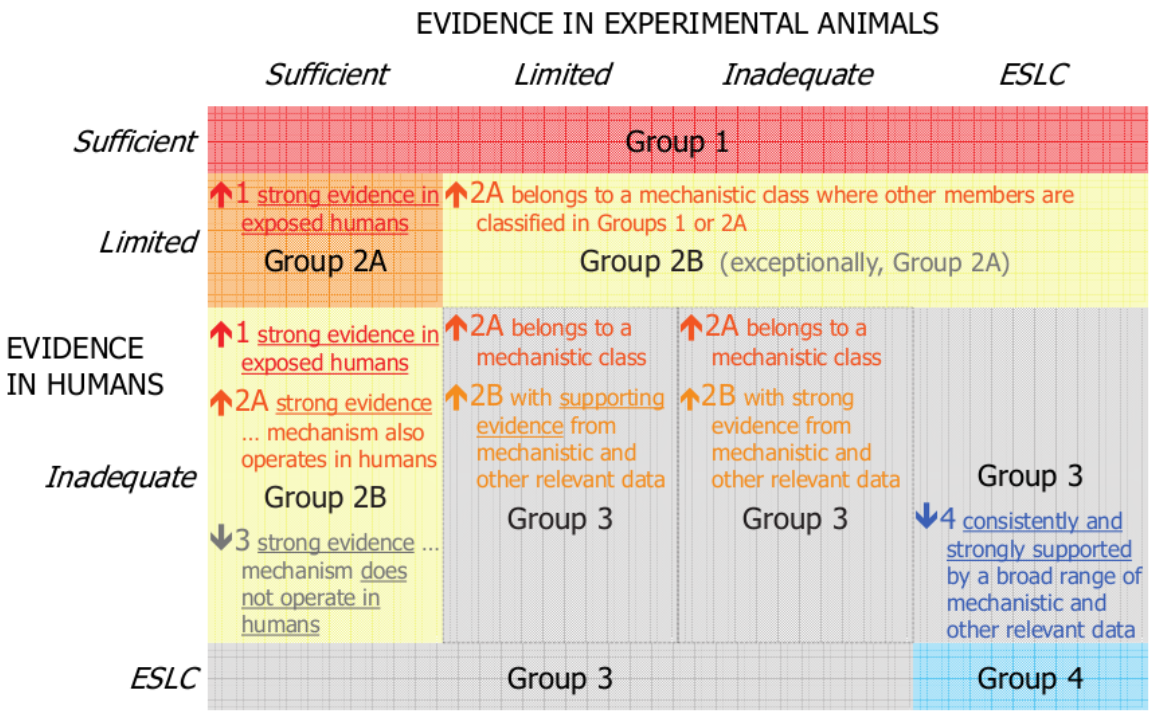
Salmon, W. C., Causality and explanation. *Oxford University Press,* **1998**

Smith, M., et al. (2016) Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environmental Health Perspectives*. 124, 6, pp.713-721.

Williamson, J., Establishing causal claims in medicine, *in press*, **2017.**

8 Appendix A. External validity and IARC

Here we compare our approach to external validity to that of the International Agency for Research on Cancer (IARC). IARC's approach is summarized in the following table (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 have a separate ranking of evidence of carcinogenicity in humans and animals:

Sufficient = established
Limited = provisionally established
Inadequate = arguable or speculative
Evidence Suggesting Lack of Carcogenicity (ESLC) = ruled out

In terms of the approach presented here, two tables are particularly relevant to the IARC categorisation. First, a table which assumes that causality in the source has been established and which 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:

Similarity of Mechanism: Causation in target pop. on the basis of evidence directly about the target:	Established	Provisionally Established	Arguable	Speculative	ruled out
Established	Established	Established	Established	Established	Established
Provisionally established	Established	Provisionally established	Provisionally established	Provisionally established	Provisionally established
Arguable	Established	Provisionally established	Arguable	Arguable	Arguable
Speculative	Established	Arguable	Speculative	Speculative	Speculative
Ruled out	Provisionally ruled out	Ruled out	Ruled out	Ruled out	Ruled out

A second table assumes 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:

Causation in source pop.: Causation in target pop. on the basis of evidence directly about the target:	Established	Provisionally Established	Arguable	Speculative	ruled out
Established	Established	Established	Established	Established	Provisionally established
Provisionally established	Established	Provisionally established	Provisionally established	Provisionally established	Arguable
Arguable	Established	Provisionally established	Arguable	Arguable	Speculative
Speculative	Established	Arguable	Speculative	Speculative	Ruled out
Ruled out	Provisionally ruled out	Ruled out	Ruled out	Ruled out	Ruled out

There is a broad agreement between the approach presented here and that of IARC. The approach presented here is simpler in one respect: a single scale from established to ruled out, rather than two different categorisations. On the other hand, our scale involves more categories and we draw a distinction between establishing causality in the source population and establishing similarity of mechanism; these differences allow for more nuance in the tables.