Chapter 6
Evaluating Evidence of Mechanisms

Abstract In this chapter, we discuss how to evaluate evidence of mechanisms. This begins with an account of how a mechanistic study provides evidence for features of specific mechanism hypotheses, laying out a three step procedure of evaluating: (1) the methods used, (2) the implementation of the methods, and (3), the stability of the results. The next step is to combine those evaluations to present the quality of evidence of the general mechanistic claim.

Having explained how evidence of mechanisms can be obtained, the next step is to evaluate that evidence, which is the topic of this chapter. In the following chapter will explain how this evaluation can be integrated with an evaluation of evidence for a correlation in order to determine an overall evaluation of the causal claim of interest.

6.1 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 Chap. 5. 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 posit key features of potential mechanisms of action; 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 when clinical

study evidence is high quality can it 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 for 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 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, 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 that 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 Fig. 3.1), one needs to combine these to come up with a final quality status for the general mechanistic claim.

The procedure of this section is summarised in Fig. 6.1. The three-step method for evaluating mechanistic studies is presented in the next section, Sect. 6.2. These steps contribute to the evaluation of the general mechanistic claim as described in Sect. 6.3. Finally, Sect. 6.4 describes how the evaluation of evidence of mechanisms can be presented.
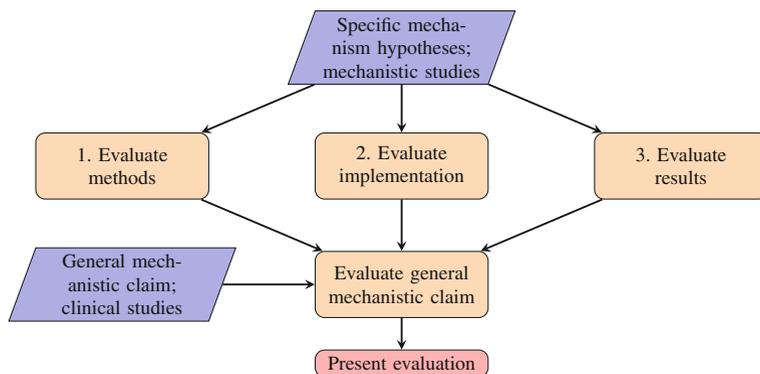

Fig. 6.1 A procedure for evaluating evidence of mechanisms

6.2 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 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 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 to 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:

4. *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:

5. *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.

6. *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. This may be understood as application of Hill's consistency indicator to evidence of mechanisms (Hill 1965).

6.3   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 mechanisms 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 (see also the tools in Chap. 4).

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

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?

1. Has a correlation of the same size been established in many studies under slightly varying circumstances (robust detectability)? If yes, is it likely that the population of interest falls within the range of circumstances which have been tested?
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?
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 or mechanisms 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 study population and the target population, and found to be similar in both, and such that it can account for the observed correlation? 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 mechanism or 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

Table 6.1 Determining the status of the general mechanistic claim (GMC) on the basis of evidence from mechanistic studies and from clinical studies

	Status of the GMC on the basis of mechanistic studies						
	Established	Provisionally established	Arguable	Speculative	Arguably false	Provisionally ruled out	Ruled out
Established	Established	Established	Established	Established	Established	Provisionally established	Speculative
Provisionally established	Established	Established	Provisionally established	Provisionally established	Arguable	Speculative	Arguably false
Arguable	Established	Provisionally established	Provisionally established	Arguable	Speculative	Arguably false	Ruled out
Speculative	Established	Provisionally established	Arguable	Speculative	Arguably false	Provisionally ruled out	Ruled out
Arguably false	Established	Arguable	Speculative	Arguably false	Provisionally ruled out	Provisionally ruled out	Ruled out
Provisionally ruled out	Arguable	Speculative	Arguably false	Provisionally ruled out	Provisionally ruled out	Ruled out	Ruled out
Ruled out	Speculative	Arguably false	Ruled out	Ruled out	Ruled out	Ruled out	Ruled out

(Row labels under "Status of GMC on basis of clinical studies")

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 6.1 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 mutually supporting roles of mechanistic studies and clinical studies. Note, finally, that determining the status of the general mechanistic claim by combining evidence from clinical and mechanistic studies should not be confused with the task of determining the status of the causal claim on the basis of the status of the general mechanistic claim and the status of the correlational claim—a point which is discussed further at the end of Sect. 7.1 when we develop the analogy of reinforced concrete.

6.4 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 rela-

tionships. 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? Questions to be considered here are, for instance (see Sects. 6.2 and 6.3): 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 (see Sect. 4.6)? 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 study 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 study?
6. What is the outcome and how is it measured in the study?
7. What is the status of the general mechanistic claim concerning similarity? Questions to be considered here are, for instance (see Sects. 6.2 and 6.3): What is the hypothesised mechanism in the study population? 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 the evidence robust?

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 Sect. 6.2. 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:

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

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 Chap. 5, and the evaluation process outlined in this section, see Sect. 4.5.

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 Sect. 4.6. Our other tools in Chap. 4 also demonstrate how the evaluation of evidence of mechanisms can be integrated into existing evidence appraisal practices.

Example: *ACE inhibitors*.

ACE inhibitors work by modulating the functioning of the 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 study and the target populations are similar enough.

However, evidence from two subgroup analyses of the ALLHAT (Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial) trial suggested that there were difficulties in establishing efficacy for ACE inhibitors in African Caribbean populations. Piller et al. (2006) showed much higher rates of angioedema (an important and serious side-effect of ACE inhibitor treatment) in African Caribbean individuals, while Leenen et al. (2006) showed that calcium channel blockers (CCB) showed better efficacy than ACEi in that population. 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 mechanistic 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 and Russo 2016).

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, in 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 a 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.

References

AHRQ (2013). Mechanistic evidence in evidence-based medicine: A conceptual framework. Technical report. Agency for Healthcare Research and Quality. https://ahrq-ehc-application.s3.amazonaws.com/media/pdf/mechanistic-evidence-framework_white-paper.pdf.

Aronson, J. (2005). Biomarkers and surrogate endpoints. *British Journal of Clinical Pharmacology*, *59*(5), 491–494.

Chow, S.-C., & Liu, J.-P. (2004). *Design and analysis of clinical trials*. Hoboken, New Jersey: Wiley.

Clarke, B., & Russo, F. (2016). Causation in medicine. In J. Marcum (Eds.), *Companion to contemporary philosophy of medicine* (pp. 297–322). Bloomsbury.

Fyhrquist, F., & Saijonmaa, O. (2008). Renin-angiotensin system revisited. *Journal of Internal Medicine*, *264*, 224–236.

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

Holman, B. (2017). Philosophers on drugs. *Synthese*. https://doi.org/10.1007/s11229-017-1642-2.

Khan, J., & Beevers, D. (2005). Management of hypertension in ethnic minorities. *Heart*, *91*(8), 1105–1109.

Leenen, F. H., Nwachuku, C. E., Black, H. R., Cushman, W. C., Davis, B. R., Simpson, L. M., et al. (2006). Clinical events in high-risk hypertensive patients randomly assigned to calcium channel blocker versus angiotensin-converting enzyme inhibitor in the antihypertensive and lipid-lowering treatment to prevent heart attack trial. *Hypertension*, *48*(3), 374–384.

Montgomery, D. (2009). *Design and analysis of experiments*. Hoboken, New Jersey: Wiley.

Piller, L. B., Ford, C. E., Davis, B. R., Nwachuku, C., Black, H. R., Oparil, S., et al. (2006). Incidence and predictors of angioedema in elderly hypertensive patients at high risk for cardiovascular disease: A report from the antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). *The Journal of Clinical Hypertension*, *8*(9), 649–658.

Rappaport, S. M., Kim, S., Lan, Q., Vermeulen, R., Waidyanatha, S., Zhang, L., et al. (2009). Evidence that humans metabolize benzene via two pathways. *Environmental Health Perspectives*, *117*(6), 946.

Thomas, R., Hubbard, A. E., McHale, C. M., Zhang, L., Rappaport, S. M., Lan, Q., et al. (2014). Characterization of changes in gene expression and biochemical pathways at low levels of benzene exposure. *PloS one*, *9*(5), e91828.