Extrapolating from model organisms in pharmacology

Veli-Pekka Parkkinen and Jon Williamson

Abstract

In this chapter we explore the process of extrapolating causal claims from model organisms to humans in pharmacology. We describe and compare four strategies of extrapolation: enumerative induction, comparative process tracing, phylogenetic reasoning, and robustness reasoning. We argue that evidence of mechanisms plays a crucial role in several strategies for extrapolation and in the underlying logic of extrapolation: the more directly a strategy establishes mechanistic similarities between a model and humans, the more reliable the extrapolation. We present case studies from the research on atherosclerosis and the development of statins, that illustrate these strategies and the role of mechanistic evidence in extrapolation.

1 Introduction

How does extrapolation work in pharmacology? In pharmacology, the causal claims of primary interest are those which assert the efficacy (or lack of efficacy) of a drug and those that assert that a drug causes a particular harm (or, alternatively, that it is safe, i.e., that it causes no significant harm). Causal claims will normally be tested on model organisms, and the results of these tests are used as evidence for corresponding causal claims on humans. This chapter aims to shed some light on this process of extrapolating causal claims from model organisms to humans in pharmacology.

The ultimate goal of pharmacological inquiry is usually to *establish* the relevant causal claim, i.e., to secure evidence which both confirms the causal claim to a sufficiently high degree and makes it sufficiently likely that further evidence will not significantly lower this degree of confirmation. Once established, a causal claim can be added to our stock of claims and treated as evidence for further claims. For instance, establishing a causal claim in a model organism allows it to be taken as evidence for the corresponding claim in humans. Establishing is very demanding, however, and lower standards of surety are sometimes also of use: a drug approval committee may only need a reasonable suspicion of harm in humans in order to reject an application for approval; if the cost of treatment is sufficiently low or the cost of failing to treat is sufficiently high, a drug may be approved for use in certain cases even where efficacy or safety in humans is not conclusively established. For example, during the 2014 Ebola outbreak, the World Health Organization recommended the use of certain yet unregistered treatments based on just the results of model organism studies, as the cost of failing to treat patients was certain to be catastrophic (World Health

Organization, 2014). Such cases are rather exceptional, however. Normally, for a new drug to be approved, it must be established to be efficacious in humans.

Exactly which claims are extrapolated from studies on model organisms depends on the stage of research and whether the causal claims in question concern efficacy or harm. At the outset of a first-in-human trial, claims of *efficacy* in humans are not yet considered established, by way of extrapolation or otherwise. At this stage the model organism results are treated as exploratory; the established efficacy in non-human animal models suggests a hypothesis about efficacy in humans. While the model organism results support this hypothesis about a corresponding effect in humans, they fall short of establishing the hypothesis. Nevertheless, significant support is required here, in order to justify the cost and risks of human trials, and extrapolation from animal studies grounds this inference.

Once trials on humans have been performed, these trials will provide the primary evidence in favour of efficacy in humans. However, extrapolation then comes into play in another way: mechanistic evidence, obtained by studies on model organisms, is used to explain and support an observed correlation between the drug and the clinical benefit. As we shall argue, the evidence obtained in model organisms remains useful at this stage because it is typically the case that the tests that can be carried out on humans are limited in key ways.

With regard to *safety and harm*, extrapolation plays an even more substantial role in the process. At the outset of a first-in-human trial, one needs to be sufficiently confident that the drug being tested has no serious negative side-effects in humans. Thus, claims about the toxicity or other side-effects of a compound need to be extrapolated from model organism studies to justify human trials.

The plan of the chapter is as follows. In §2 we discuss some of the challenges that face extrapolation in pharmacology. §3 provides some examples of extrapolation in pharmacology that inform the rest of the chapter. In §4 we present four strategies for extrapolation and in §5 we see—by appealing to a thesis concerning the role of evidence of mechanisms in establishing causal claims in the biomedical sciences—why these strategies work when they do work. In §6 we argue that this analysis supports a recent movement to evaluate evidence of mechanisms in a more rigorous way in the biomedical sciences, and we discuss a recent objection to mechanism-based extrapolation.

2 Model organisms in pharmacology

Model organisms as diverse as yeast, rats, and non-human primates are used extensively in pharmacological research, both for identifying potential targets for drugs, and for safety and efficacy testing. In the first case, the purpose of studying a model is to identify a component of a pathophysiological mechanism as a potential target for intervention. In the latter case the task is to test whether intervening on a mechanism in a particular way produces desired outcomes. In both cases, one must deal with the uncertainty inherent in transferring the results from the model organisms to humans.

Evaluating the evidence from model organism studies relies on judgements about relevant similarities between the model and humans. These considerations differ slightly depending on the intended purpose of the model study. In target

identification, one attempts to establish the biochemical properties of a component entity that would allow pharmacological interventions to be targeted on it. The required similarity here involves parts of a mechanism—model studies suffice to establish an entry point into a mechanism as long as the component(s) of interest are similar between humans and the model organisms in both structure and function, even if other parts of the mechanism are dissimilar. Component similarities are commonly established by appeal to shared evolutionary ancestry that can establish homology, a strategy we describe and scrutinize below in §4. Note that relevant similarity here involves a particularly robust sense of functional homology, i.e., not just similarity of selected function, nor similarity of structure (Love, 2007). For target identification or the study of molecular level effects of a drug, it is often preferable to work with a model in which the mechanism of interest is well understood, even if there is reason to believe that the mechanism as a whole is in some ways dissimilar in humans. Thus, models as distant to humans as yeast are used when the primary focus is discovering novel targets for pharmacological interventions or detecting fine grained effects.

In safety and efficacy testing the problem of extrapolation is slightly different and arguably more difficult. Here the model-based inferences are not about properties of individual components of a mechanism, but about the effects that interventions on the mechanism cause to the well-being of the organism as a whole. To infer an organism-level outcome in humans based on results obtained in a model organism, one would need to establish that the output of the whole mechanism under interventions is similar in humans despite possible differences in some of the mechanism's parts or its causal environment. This cannot usually be assumed, even if the model and human mechanisms are known to be structurally similar in parts and the mechanism is well understood in the model. Biological mechanisms are typically complex in the sense that many component functions are coupled—connections between components are abundant and exhibit feedback—such that even a detailed understanding of the component properties will not allow deriving specific predictions about what would happen to the output of the mechanism under interventions if some individual components were changed. Nor does establishing structural similarities guarantee similarity of function. In addition, unexpected interactions between the mechanism and its environment might modulate or mask the typical output of the mechanism in humans, and these effects cannot be predicted just by knowing that humans exhibit a similar mechanism to one that is well understood in the model. A tragic example of translational failure despite established similarity of the mechanism of action is the TGN1412 trial, in which an administration of an immunomodulatory drug whose target receptor was well characterized resulted in catastrophic conditions in humans, despite being proven safe in trials on monkeys that share the same receptor with minimal structural differences (Kenter and Cohen, 2006).

TGN1412 is a humanised antibody that is a strong agonist of the CD28 receptor on human T cells, a type of white blood cell that is a part of cell-mediated immune system. TGN1412 is capable of activating T cells irrespective of the presence of other regulatory signals typically required for T cell activation. This capacity promised great therapeutic potential, as T cell regulation can play a role in the treatment of many autoimmune diseases and cancer. Before the human trial, TGN1412 was tested on cynomolgus and rhesus monkeys. These were considered valid models of humans due to perfect sequence homology of the extracellular domain of CD28—the part of the receptor that lies outside the cell

and binds signalling molecules (Attarwala, 2010). Based on these studies, a no-observed adverse effect level (NOAEL) was determined. In the first-in-human trial, six human volunteers were administered a dose that was one five hundredth of the estimated NOAEL. The result of the trial was a tragedy: one of the participants died, and the rest suffered possibly irreversible adverse effects due to a rare immune system response called cytokine storm. Cytokines are signalling factors that normally play an adaptive role in the immune system, activating immune cells to attack invasive pathogens and to produce more cytokines. In a cytokine storm, this feedback loop runs out of control, causing local activation of too many immune cells and subsequent damage to any affected organs. Nothing like such a reaction had been seen in the experiments on monkeys. It has been suggested that the drastic difference between the monkey and human outcomes is explained by the fact that the monkey equivalent of the cell type that drove the cytokine storm in humans lacks CD28 receptors, and was thus not activated by TGN1412 (Eastwood et al., 2010). To summarize, even though humans share TGN1412's mechanism of action as it operates in monkeys, humans have in addition other mechanisms sensitive to TGN1412, which were responsible the effects not seen in monkeys.

Extrapolating causal effects of pharmacological interventions from model organisms to humans is thus risky even when one has knowledge of some relevant mechanistic similarities. What one can do to alleviate this risk is to search for outcomes that are robust against changes in background conditions or parts of the supporting mechanism. This involves testing an intervention in a range of models that differ from each other as well as from humans, and searching for convergent results across the various model studies. If it can be shown that an outcome is independent of physiological features idiosyncratic to any particular model, the strict assumption about similarity between models and humans can be relaxed. This strategy is described and analysed in more detail in §4.

3 Case studies

This section describes examples of model organism research in the discovery and development of statins for treating and preventing heart disease. Statins reduce the level of cholesterol in blood, thus reducing the risk of atherosclerosis and subsequent heart disease. Statins produce their effect by inhibiting the activity of HMG-CoA reductase (HMGCR), a rate-limiting enzyme in cholesterol biosynthesis. HMG-CoA reductase catalyses a reaction in which HMG-CoA is converted into mevalonic acid, a precursor of cholesterol. Statins mimic the structure of HMGCR and compete with it in binding HMG-CoA, but do not have the same enzymatic function. This reduces the rate at which mevalonic acid is produced and thus controls the rate of cholesterol synthesis in the liver. Statins are now widely used as preventive treatment for heart disease, and their efficacy has been demonstrated in large randomized clinical trials.

The focus on cholesterol in the treatment of heart disease is based on decades of experimental and epidemiological research on the connection between elevated blood cholesterol and cardiovascular events (Steinberg, 2007). The earliest evidence suggesting that this connection is causal came from experiments on the effects of cholesterol feeding in rabbits, which indicated that high blood cholesterol level is linked to atherosclerosis – the thickening and hardening of the

artery wall in a manner that occludes blood flow, causing the cardiac events that characterise coronary heart disease (Anitschkow, 1913). These results were subsequently replicated in other species with certain exceptions: some model species such as rats and dogs failed to show similar susceptibility to cholesterol induced atherosclerosis (Bruger and Oppenheim, 1951). At the time of these early experiments, the researchers lacked detailed knowledge of the pathophysiology of atherosclerosis, and therefore could not validate animal models by directly comparing the relevant mechanisms to humans. But once the number of successful experiments in many different model species grew, one could argue that cholesterol's atherogenic potential had been shown to be independent of the specific physiology of any particular model to such a degree that it warrants inferring the existence of a mechanism extrapolatable to humans (Parkkinen, 2016). Such reasoning appeals to the robustness of evidence, a strategy we describe in the next section.

It is only after the experimental animal results had established a manipulable link between cholesterol and atherosclerosis that large-scale epidemiological studies on the link between cholesterol and heart disease were conducted (Steinberg, 2007, pp. 33-39). Given experimental evidence from model organism studies, and epidemiological evidence of correlation between blood cholesterol and heart disease in humans, the 'cholesterol conception of atherosclerosis' was established as an explanation of the prevalence of atherosclerotic heart disease in populations characterized by high lipid consumption and high average blood cholesterol levels. This sparked an interest in the details of cholesterol's role in the pathophysiology of atherosclerosis and a search for effective cholesterol lowering drugs.

The first statin, known today as compactin, was isolated by Akira Endo and his collaborators at Sankyo Research Laboratories in 1972 (Endo, 2010, p. 487). Compactin was shown to be a highly efficient inhibitor of HMGCR in mammalian cell cultures (Endo, 2010, pp. 487-488). The first studies testing the efficacy of compactin for cholesterol lowering were carried out in rats. These studies, somewhat surprisingly, showed virtually no effect on blood cholesterol levels (Endo, 2010, p. 488). The failure of the first animal tests led Sankyo to effectively drop compactin from the drug development pipeline, but Endo was allowed to carry on studying its mechanism of action. This research led to a hypothesis that could explain the results seen in rats, while suggesting that compactin would be efficacious in many other species including humans. The working hypothesis had been that inhibiting cholesterol metabolism in the liver would lead to a reduction in serum cholesterol due to an increase in the extraction of cholesterol from plasma lipoproteins to support normal cellular functions. What Endo and his collaborators discovered is that this is not what happens in rats: the rat liver is incapable of catabolizing lipoproteins. Instead, the competition between compactin and HMGCR was compensated by upregulating HMGCR in the liver, directly cancelling the effect of compactin (Endo et al., 1979). This suggested that compactin would be efficacious in species that do not exhibit similar regulation of HMGCR. Importantly, a sub-population of humans suffering from hypercholesterolemia completely lack this regulatory mechanism due to a genetic dysfunction. Subsequent experiments showed that this explanation is likely to be true: compactin was shown to have a significant cholesterol lowering effect in other animal models such as dogs and monkeys. The important point to note for our purposes is that the rat model that was initially used

was not validated by either direct comparison of the relevant mechanisms to humans, or by auxiliary evidence to justify an assumption about similarity of mechanisms. Once Endo and collaborators came up with a testable mechanistic explanation for the results, it became apparent that the rat is not a valid model of humans for testing the efficacy of statin treatment, due to differences in feedback regulation of cholesterol biosynthesis.

The examples discussed above consider a use of model organisms where some outcome variable of clinical interest, e.g., blood cholesterol level, is measurable both in the model and in humans, so that the model results can be taken to represent the behaviour of a corresponding variable in humans. But model organisms need not express an outcome variable similar to humans to be useful, if they nonetheless host similar mechanisms. As an example, consider a study by Maciejak et al. (2013), investigating the molecular and cellular level effects of statins using the yeast Saccharomyces cerevisiae as a model. S. cerevisiae is one of the most extensively studied eukaroytes due to its short generation time, low-maintenance culturability, and well-understood genetic architecture that amends itself to manipulation. What makes it a feasible model for pharmacological research is the fact that despite remarkable evolutionary distance, many biochemical pathways are at least partly conserved between yeast and humans. This is the case with the sterol pathway responsible for cholesterol synthesis in humans: S. cerevisiae hosts two homologs of the human HMGCR coding gene, and the pathway is biochemically similar to the human version with the exception that the end product synthesized in yeast is ergosterol instead of cholesterol. Maciejak et al. studied the effects of four different statins on the behaviour of the sterol pathway and cell growth in three yeast models: two models that carried either one of the native yeast HMGCR coding genes, and one engineered to express the human version of the gene (Maciejak et al., 2013). This allowed them to draw a number of conclusions about the potency of each statin, and the number and severity of side effects on other cellular processes. Understanding these cellular level effects is important from a clinical point of view, as many other cellular functions are dependent on the sterol pathway. The most general conclusion was that statin treatment will trigger compensatory upregulation of many sterol and non-sterol pathway genes, a result that could be validated by comparison to human cell cultures. These results are informative about the clinical efficacy of statins even though the model system does not express anything like the clinical variable of interest, as they provide clues to how the underlying mechanism can be manipulated without producing unwanted side-effects. In this case, extrapolating the model results to humans rests on knowledge of evolutionary conservation of the relevant mechanisms. Note in addition that, as explained in §1, precautionary reasoning suggests that the threshold of evidence required for establishing extrapolation should be considered lower in this case, as the inferences drawn based on the model study consider possible harmful side-effects rather than efficacy.

4 Strategies for extrapolation

In this section we present four general strategies for extrapolating from animals to humans: enumerative induction, comparative process tracing, phylogenetic reasoning, and robustness analysis.

Enumerative induction. The most straightforward way to extrapolate from animal models to humans is to collect evidence from many models and generalize the results to humans by simple enumerative induction. This strategy makes no appeal to evidence of mechanisms. It is simply the similarity of the observed phenomenon in many non-human instances that supports the inference to the human case. This strategy is risky: the fact that many non-human animals or in vitro models share some feature or respond similarly to an intervention is no guarantee that humans will respond the same way. To justify extrapolation, it is typically better to have some further evidence of similarity of the relevant mechanisms. In an ideal case, one knows the details of a relevant mechanism both in the model and in humans, and can establish the similarity of the mechanisms by direct comparison.

Comparative process tracing. In reality, one is rarely in a position where a model is verifiably identical to humans with respect to the relevant mechanisms. Instead, biomedical scientists typically argue for the validity of their models based on partial mechanistic similarity between the models and humans. For partial mechanistic similarities to support extrapolation, one needs a method for investigating how differences in parts of a mechanism might result in differences in the mechanism's output. Daniel Steel calls such a method comparative process tracing, and argues that in many cases this method will be able to justify extrapolation (Steel, 2008). Comparative process tracing starts with a search for component activities that act as bottlenecks for causal influence within the mechanism, such that, once the behaviour of the bottleneck is held fixed, the behaviour of any components causally downstream from it is not influenced by components causally upstream to it. If one is able to identify such critical components, one only needs to establish that the components between the critical bottleneck and the endpoint of interest are similar in the model and humans; establishing full mechanistic similarity is not necessary for extrapolation to be reliable (Steel, 2008, p. 89).

But direct evidence of even partial similarity of mechanisms is not always necessary; there are other strategies to establish similarity of mechanisms. Below we describe two such strategies: phylogenetic reasoning and robustness analysis. Despite differences in how these strategies are implemented in research, there are similarities in the rationale that underlies their use. Both strategies work by making an explanation of the behaviour of the model in terms of mechanisms idiosyncratic to the model less likely than an explanation that posits similarities between the model and humans in the relevant mechanisms.

Phylogenetic reasoning. Choice of a model organism is often guided by evolutionary considerations: extrapolation from a model organism to humans is considered more secure if the relevant mechanism is conserved between the model and humans, than if the mechanisms have evolved independently. Directly establishing that a mechanism is conserved, however, would amount to directly observing that the mechanisms are similar—no specifically evolutionary reasoning would be needed to justify the extrapolation. But phylogenetic information can and is being employed in model organism research in a different way, to establish assumptions about similarity of mechanisms in the absence of detailed knowledge of the mechanism in the target of extrapolation. This

involves a two-step inference, where each step relies on supplemental empirical evidence other than direct comparisons between the model and the target (Levy and Currie, 2015). In the first step, one observes the trait of interest in the model and other closely related species to establish that the trait is widely shared among members of a clade. This evidence is used to project the trait to an ancestor species based on the assumption that a trait shared by many members of a clade has likely evolved from a similar trait in their common ancestor. In the second step, the trait is projected to the target species based on empirical evidence that the target species, too, shares a common ancestor with the model.

The inference from the model to the target relies on knowledge of the nature of the evolutionary process. The fact—established by supplementing phylogenetic evidence—that the trait of interest has evolved from the same ancestral trait constrains the amount of variation between the model and the target, as any evolutionary novelty in the trait must be achievable by piecemeal modification of the ancestral form and must result in a viable phenotype. Given this fact about evolution, the reasoning from the model to the target can be reconstructed as an argument for the truth of a common descent explanation that, if true, would imply trait similarities between them.

Common descent explains the similarity of traits between species by constraining the possible variation that the species can exhibit in the trait. The explanation has a contrastive structure: the explanans (common descent) favours the explanandum (trait similarity) at the expense of an exclusive alternative that would be the case if the explanans were false, i.e., had the trait of interest evolved independently in the two species, it would exhibit more (or be more likely to exhibit more) variation between the species than is actually the case. Thus, if a common descent explanation applies to the trait of interest in the model and target species, they are likely to be similar. A phylogenetic argument aims to show that the common descent explanation is more likely to be true of the trait of interest than an explanation that posits independent evolution of the trait in the two species. The first step is an overtly abductive inference: common descent would be the best explanation of trait similarities observed between the model and other species in the clade. In the second step this explanation is applied to the target species based on supplemental phylogenetic evidence that the target species shares a common ancestor with the model.

Some elaboration is in order concerning what can be learned about causality by means of this strategy. Phylogenetic arguments work best for establishing similarities in components of a mechanism. However, more than mechanistic similarity between a model organism and humans is needed to establish causal claims about the mechanism's output in humans, even if the mechanism's behaviour under experimental interventions is well characterized in the model. This is because of the problem of masking. The problem of masking refers to a situation in which there are two or more separate mechanisms linking a cause variable to some effect variable of interest, such that the influence of these mechanisms on the effect differs. The mechanisms might for instance partly or completely cancel out, or modulate each other's output in a non-additive manner. In such a case, establishing just one of the mechanisms, or even establishing the behaviour of all of them independent of each other, will not support reliable inferences about the behaviour of the effect under interventions on the cause.

Thus, establishing similarity of a mechanism between an animal model and humans will not support the extrapolation of an organism-level causal effect from the model to humans, unless one can in addition rule out that humans exhibit masking mechanisms that are not present in the animal model. The TGN1412 catastrophe mentioned earlier is an example of how difficult this can be. Even though humans share the mechanism responsible for the results that were seen in the models, this alone was not enough to predict the outcome of the experiment in humans, as humans exhibit in addition other mechanisms sensitive to TGN1412, and the operation of these mechanisms causes the human immune system to react drastically differently. For establishing similarities in the output of a mechanism under interventions on its components, one would need to establish that the operation of the mechanism is insensitive to differences not only in some of its parts, but also the causal context in which it is embedded, which may include other mechanisms relevant for the outcome of interest. Phylogenetic reasoning can rarely establish this.

Robustness analysis. Extrapolating organism-level outcomes of pharmacological interventions, such as the results of toxicological tests, rests on the claim that the mechanism responsible for the effect is not too dissimilar between the model and humans, and that there are no interfering mechanisms in humans that could mask or modulate the effect. These assumptions can sometimes be partly justified by phylogenetic arguments or by deliberately engineering the model to make it similar to humans, but typically there is considerable uncertainty about the representativeness of model results with respect to humans. Moreover, it is often unclear how much detail about the mechanism and its environment one would need to know in order to know that the extrapolation is reliable. What one can do is to try to weaken or discharge the underlying assumptions by demonstrating that similar results can be produced in many different model organisms that vary in relevant parts of their physiology, thus demonstrating that the outcome is independent of the idiosyncratic features of any particular model. When each model organism is treated as a source of evidence—fallible due to causal dissimilarity that introduces error in the extrapolation to humans—the reasoning can be seen as a kind of robustness analysis.

Wimsatt (2007) describes robustness analysis as consisting of four procedures:

> 1. To analyse a variety of independent derivation, identification, or measurement processes. 2. To look for and analyse things that are invariant over or identical in the conclusions or results of these processes. 3. To determine the scope of the processes across which they are invariant and the conditions on which their invariance depends. 4. To analyse and explain any relevant failures of invariance. I call things that are invariant under this analysis robust (Wimsatt, 2007, p. 44).

Wimsatt's procedures are meant to capture a highly general reasoning strategy that is used to study the reliability of evidence and scientific inference; the processes to which robustness analysis is applied can be anything from sensory modalities to derivations of results from mathematical models (Wimsatt, 2007, pp. 45-46). Here we consider *empirical robustness analysis*, where the processes

studied are experimental procedures and the results obtained through them. Empirical robustness analysis can be reconstructed as explanatory reasoning that works by ruling out explanations of experimental results in terms of idiosyncrasies of the causal set-up of any particular experiment (cf. Schupbach, 2016). Every experiment or detection method has its own sources of error that make the evidence obtained from it fallible: instead of reliably tracking a phenomenon existing independently in nature, our experimental results might be artefacts created by the methods themselves. However, it would be unlikely that many methods based on different causal principles would produce similar artefacts. As an illustration, consider two experimental set-ups, one that exhibits error source E, and one that exhibits error source E* that is independent of E. Often, if both of these experiments produce a concordant result, one can rule out that the result is an effect of error source E, on the grounds that it is unlikely that both experiments would be independently erroneous but nonetheless yield the same result. By collecting evidence from many different experiments that exhibit independent sources of error, one can then infer that the best explanation for any concordant result is that the result is caused by a real underlying mechanism, rather than being a side product of the causal set-up of the experimental procedures themselves.

Similar reasoning can be applied to the evaluation of model organism studies as evidence. Extrapolation from a particular model organism rests on the assumption that the model and humans are similar with respect to relevant mechanisms. But all model organisms are dissimilar to humans in some ways, and these dissimilarities constitute a potential source of error when the model result is taken as evidence for inferences about humans. Therefore, one should initially have low credence in conclusions one draws based on model studies, unless one has direct evidence of relevant mechanisms to rule out the possibility of error. However, even if all model organisms are causally dissimilar to humans, the same evolutionary processes responsible for their dissimilarity to humans make different model organisms dissimilar to each other as well. Model organisms that are not close phylogenetic relatives are thus likely to exhibit independent errors with respect to extrapolation of the results to humans. This fact makes it possible to apply robustness reasoning to the evaluation of model organism evidence: a result that can be reproduced across heterogeneous pool of model species is independent of the idiosyncratic causal make-up of any particular model.

Under favourable conditions, robustness analysis can complement more directly attainable mechanistic evidence in extrapolation of causal claims from animal models to humans. Transferring a causal claim based on mechanistic evidence alone is susceptible to masking; not only differences within the underlying mechanism, but also differences in its causal context can be error sources that may defeat extrapolation. But searching for robust results in a pool of models where the mechanism itself as well as its physiological context varies sufficiently can help with the masking problem. Here the important aspect of robustness analysis is Wimsatt's procedure 4., the analysis and explanation of discordant results. Any failures of robustness are suggestive of either species differences in the hypothesized mechanism of action, or of the presence of masking mechanisms in some species. Analysing and explaining failures of robustness by searching for masking mechanisms in the models can help evaluate the reliability of extrapolation by guiding the search of relevant masking factors in humans.

It should be noted that this requires some explicit comparisons of mechanistic detail between the models and humans: simply seeing robust results in heterogeneous models, without knowing anything about the relevant mechanisms, does not rule out the possibility of masking in humans.

5 The logic of extrapolation

In this section we shall show how one recent line of work on the epistemology of causality can shed some light on the logic of the sort of extrapolations exemplified by the pharmacological case studies of §3.

Russo and Williamson (2007) argued that, in order to establish a causal claim, one normally needs to establish both that the putative cause and effect are appropriately correlated and that there is some underlying mechanism by which one can explain instances of the putative effect in terms of the putative cause and which can account for the observed correlation. It is not sufficient just to establish the existence of an appropriate correlation, because some correlations are not causal—they may be attributable to confounding, bias, chance or some non-causal connection between the variables of interest. Establishing the existence of an appropriate mechanism rules out these other explanations of the correlation. On the other hand, it is not sufficient just to establish the existence of an appropriate mechanism, because there may be counteracting mechanisms which cancel out or reverse the influence from the putative cause to the putative effect—this is the problem of masking. Establishing an appropriate correlation rules out such cases.

Note that the thesis is that one *normally* needs to establish an appropriate correlation and an appropriate mechanism. Only 'normally' because there are some awkward cases. For example, there are cases of causation without a linking mechanism. If the putative cause and / or the effect is an absence—something *not* happening, or a quantity that is absent—then there can be no physical mechanism linking the two. The speaker failing to catch a flight to a conference causes the absence of her talk, even though there is no physical connection between the two. In such cases we attribute causation when we can expect a mechanism linking counterfactual presences: if she were to have caught the flight, that flight and connecting transport would have provided the mechanism that gets her to the auditorium to give the talk. Moreover, there are cases of causation without a correlation: if the ball can descend to the bottom of the pinball machine whichever route it takes, then taking one particular route causes it to get to the bottom even though it does not increase the chances of it getting to the bottom. In such cases we attribute causation when we can expect a correlation were we to counterfactually block one or more of the alternative pathways. There are even cases of causation without either a correlation or a mechanism, as Longworth (2006, §4.1) points out. Such cases show that the epistemological thesis stated above is a first approximation to a more nuanced thesis. Fortunately, these nuances will not be important in what follows and we may stick with the above formulation.

The epistemological thesis has generated some controversy (see, e.g., Weber, 2007, 2009; Campaner, 2011; Clarke, 2011; Darby and Williamson, 2011; Gillies, 2011; Illari, 2011; Howick, 2011a,b; Russo and Williamson, 2011a,b; Campaner and Galavotti, 2012; Claveau, 2012; Dragulinescu, 2012; Clarke et al., 2013,

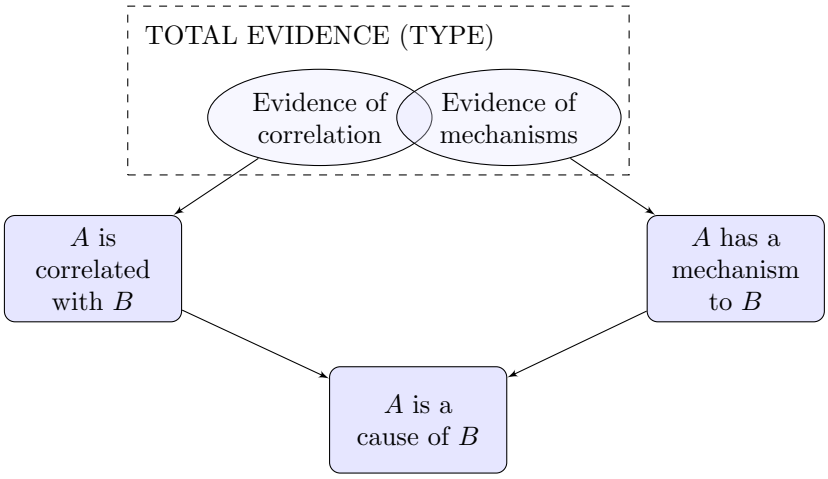

Figure 1: The epistemology of causality motivated by Russo and Williamson (2007).

2014; Fiorentino and Dammann, 2015). However, we shall neither offer a detailed justification nor a defence of the thesis here. Instead we shall assume that the thesis is correct and show that it can shed any light on the process of extrapolation.

The epistemological thesis leads to the picture of Fig. 1: some of the total available evidence is evidence for or against the existence of a correlation; some of it is evidence for or against the existence of a suitable mechanism; these both provide evidence for or against causation. Note that some items of evidence can be both relevant to correlation and mechanism. Arguably, in the right conditions a large, well-conducted randomised controlled trial (RCT) can provide evidence both of correlation and that this correlation is not spurious, i.e., that there must be some suitable underlying mechanism that accounts for the correlation. It is very rare, however, for a single study on its own to *establish* both correlation and mechanism. More often, while a large enough correlation in the sample might be enough to establish a corresponding correlation in the population, significant doubts about mechanism will remain and other evidence of mechanisms needs to be obtained or invoked in order to establish mechanism and thereby establish causality. Fig. 2 provides an alternative view of the epistemology of causality, splitting the relevant evidence into statistical trials—such as RCTs and observational studies—and other relevant evidence—such as evidence obtained from biomedical imaging, in vitro experimentation, and simulation.

The epistemological thesis also motivates a particular view as to the logic of extrapolation. This is depicted in Fig. 3. In this diagram, dashed arrows represent weak evidential relationships and full arrows represent strong evidential relationships. Normally, one can conduct more conclusive trials on a model organism population than on a human population—this is why model organisms are so important. For example, one can often perform trials on a model organism that, for ethical reasons, cannot be undertaken in humans. These trials can be better randomised in model organisms than in humans, and can

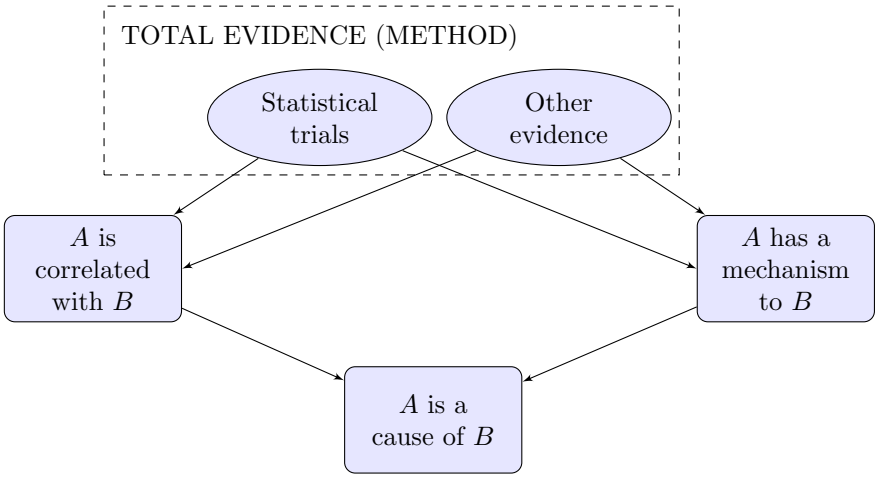

Figure 2: An alternative view of the epistemology of causality.

involve more invasive measurements, creating more conclusive evidence about causality in the model organism than what corresponding trials in humans could establish about humans. The obvious downside is that the evidential relevance of the model organism results to claims about humans may be uncertain. Thus, Fig. 3 depicts strong evidential relationships from statistical trials in animals, but weak connections in humans. Given these weak connections, it can be hard to establish a suitable correlation in humans, and thereby hard to establish causality in humans. Although studies in humans may suggest a correlation, this observed correlation may be spurious—e.g., due to confounding, bias, or chance. However, if one can establish causality in the model organism and one can establish that the mechanisms which underpin this causal relationship are sufficiently similar to those in humans, then this lends further credence to the claim that the correlation observed in humans is not spurious, i.e., that there is a genuine correlation in the underlying population. Although each consideration on its own provides rather weak evidence of correlation in humans, the combination of all these factors can, in the the right circumstances, establish correlation, and thereby help to establish causality in humans.

Let us revisit the statin case studies of §3. Recall that compactin was first tested on rats. Here, the trial failed to establish a correlation. Thus, the model organism clearly could not be used as the basis for an inference to the effectiveness of compactin in humans, even in the weak sense of motivating a hypothesis that compactin would be effective in humans. Moreover, other evidence was inadequate to establish that the underlying mechanisms in rats and humans were sufficiently similar. Hence, the model organism could also not be used as the basis for an inference to the lack of effectiveness of compactin in humans, although it did provide some evidence against a correlation in humans. Further research on the mechanisms involved led to trials on dogs and monkeys. These trials did establish a correlation in the model organisms, and, at this stage, there was other evidence of both an appropriate mechanism of action in the model organisms and of a similarity in mechanisms between the model organ-

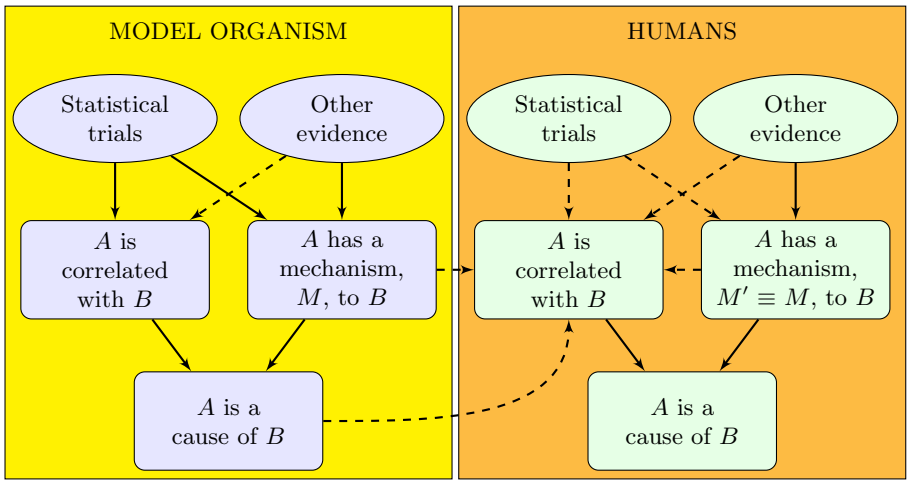

Figure 3: The logic of extrapolation as motivated by the epistemological thesis.

isms and humans. Hence the model organisms could now be used as a basis for extrapolating effectiveness to humans, at least in the weak sense of raising one's confidence in a hypothesis about effectiveness in humans to a degree that warrants conducting a trial in humans.

Let us turn to the case study involving the yeast Saccharomyces cerevisiae. Yeast is not an animal and many of its mechanisms differ from those in humans; nevertheless, extrapolation from yeast to humans follows the pattern depicted in Fig. 3. In this case the causal claims in question are rather lower-level, involving the effect of statins on the upregulation of particular genes. There is sufficient similarity between yeast and humans to extrapolate certain of these low level causal claims from the yeast to humans, however, it should be noted that no claim about efficacy with respect to clinical endpoints of interest in humans is extrapolated here. Furthermore, the low-level claims themselves provide evidence for higher level claims of the efficacy of statins for lowering blood cholesterol in humans: they help to establish the existence of appropriate mechanisms of action of statins for lowering cholesterol, and they help to show that these mechanisms of action are shared between humans and model organisms such as monkeys. The reasoning behind this last inference to similarity of mechanism of action between humans and model organisms is typically phylogenetic (§4). Certain mechanisms are shared between yeast, monkeys, dogs and humans, despite evolutionary distance, because of common ancestry. Although experiments in rats undermine robustness of efficacy of statins across species, these results can be explained away, again by invoking evidence of mechanisms.

6 Conclusion

According to the above analysis, problems of extrapolation from model organisms to humans hinge upon evidence of mechanisms. Extrapolation works by establishing causation in the model organism and establishing similarity of the model organism to humans. Establishing causation in the model organism re-

quires establishing the existence, if not the nature, of some suitable mechanism of action. The similarity that needs to be established is similarity of the mechanisms of action in the model organism to those in humans.

The strategies for extrapolation that lend most surety to the causal conclusion in humans are those that most directly establish similarity of mechanisms: comparative process tracing, phylogenetic reasoning, and robustness analysis. The first strategy employs evidence of partial similarities in mechanisms, and a search for crucial causal bottlenecks to limit the number of comparisons between a model and humans required to establish the validity of the model. The next two strategies involve an abductive inference as a key ingredient; both strategies work by ruling out explanations of model results in terms of mechanisms idiosyncratic to the model, and in favour of explanations in terms of similar mechanisms. Enumerative induction proceeds rather differently from the above three strategies: reasoning from the causal claim having been found to hold in previously observed species to the claim holding in humans. As a subsequent inference, one might also infer similarity of mechanisms as the best explanation of the causal claim holding across all the species under considerations. Similarity of mechanisms is thus inferred only indirectly, by chaining a simple induction and an inference to the best explanation. This form of inference is error-prone, if not entirely tenuous.

Given the importance of evidence of mechanisms for successful extrapolation, it becomes equally important to ascertain the quality of such evidence. Grading quality of mechanistic evidence is important for several reasons: in order to determine how credible are the mechanistic claims that the evidence supports (in particular, whether they can be considered *established* by the evidence); in order to avoid erroneous and fallacious mechanistic inferences, where possible; and in order to decide when more evidence is needed (when to commission further research). These considerations motivate the EBM+ approach to evaluating evidence in medicine (ebmplus.org), which seeks to evaluate the full range of mechanistic evidence alongside evidence of correlation, instead of focusing exclusively or almost exclusively on statistical studies, as is common in current EBM practice (Clarke et al., 2014; Parkkinen et al., 2018).

It is important to note that the evidential requirements of extrapolation vary depending on the level and specificity of the causal claim in question. Low level causal claims about interactions between individual components do not require establishing detailed similarities of the whole embedding mechanism. For example, if one is merely interested in the binding affinity of a compound to its intended molecular target, one clearly does not need a model that resembles humans with respect to the whole complex mechanism of which the target is a component of in humans. Thus, even model organisms that are phenotypically very dissimilar to humans may serve as fairly reliable sources of extrapolation, such as in the case of using yeast as a model for studying molecular level side-effects of statins. By contrast, extrapolating claims about phenotypically high level effects requires somewhat better understanding of the behaviour of the underlying mechanism as a whole, and more evidence of similarity of the mechanism as well as its causal context between the model and humans, as extrapolating such effects is susceptible to the problem of masking. Similarly, qualitative claims tolerate more differences between the model mechanisms and the human mechanisms than claims about specific quantitative effects of interventions. For example, inferences about causal relations where the exact timing

of the effect in response to the intervention is of clinical interest require higher quality evidence of mechanistic similarities than extrapolation of claims about the mere qualitative effects of the intervention. In many cases, evidence of similarity of mechanisms can be rather indirect, involving functional similarity or confirmed theory, rather than detailed knowledge of bottlenecks and other components of mechanisms, as required by comparative process tracing (Guala, 2010). Thus, the strategies of §4 should be thought of as examples of ways of generating evidence of similarity of mechanisms, but not exhaustive.

One prominent criticism of mechanism-based extrapolation overshoots its conclusions because it neglects the nuances mentioned above. Howick et al. (2013a,b) have argued that mechanistic evidence typically fails to support extrapolation in the intended way, and this is due to the inherent complexity of biological mechanisms, unexpected interactions between many mechanisms, and the uncertainty concerning how well the mechanism must be understood for one to know whether extrapolation is justified or not. They argue that predicting clinically relevant outcomes from mechanistic knowledge is often fallible, and then argue that this inherent uncertainty compromises most attempts to extrapolate from one context to another based on mechanistic knowledge (Howick et al., 2013b, pp. 281-285). We agree that one often cannot reliably predict outcomes of interventions from knowledge of mechanisms alone—evidence of correlation is also required (§5). The pessimistic conclusion of Howick et al. concerning extrapolation in general does not follow from this. In the pharmacological cases we have discussed, one has evidence of some causal effect of an intervention in a model already; as explained in §5, this is not inferred solely from knowledge of a mechanism, and indeed the details of the mechanism need not be known at all. The extrapolation task considers the use of mechanistic evidence to evaluate whether similar effects would be seen in humans. Howick et al.'s worries about complexity and contextual effects do apply to such inferences, but not with similar force in every case. As we have explained, it is the extrapolation of complex, whole organism level outcomes that is most susceptible to error due to contextual masking factors and insufficient mechanistic evidence, but even these problems can be mitigated by testing the outcome for robustness in a heterogeneous pool of models. When it comes to extrapolation of lower level effects, the evidential requirements of extrapolation are less demanding and more clearly delineated.

Acknowledgements. This research was carried out as a part of the projects *Grading evidence of mechanisms in physics and biology*, supported by the Leverhulme Trust, and *Evaluating evidence in medicine*, supported by the UK Arts and Humanities Research Council. We are grateful to the referees and editors for helpful comments.

References

Anitschkow, N. (1913). Über die Veränderungen der Kanichnenaorta bei experimenteller Cholesterinsteatose. *Beitrage zur pathologischen Anatomie und zur allgemeinen Pathologie*, 56:379–404.

Attarwala, H. (2010). TGN1412: From discovery to disaster. *Journal of Young Pharmacists*, 2(3):332 – 336.

Bruger, M. and Oppenheim, E. (1951). Experimental and human atherosclerosis: Possible relationship and present status. *Bulletin of the New York Academy of Medicine*, 27(9):536.

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

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

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

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

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

Claveau, F. (2012). The Russo-Williamson theses in the social sciences: Causal inference drawing on two types of evidence. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4):806–813.

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

Dragulinescu, S. (2012). On 'stabilising' medical mechanisms, truth-makers and epistemic causality: a critique to Williamson and Russo's approach. *Synthese*, 187(2):785–800.

Eastwood, D., Findlay, L., Poole, S., Bird, C., Wadhwa, M., Moore, M., Burns, C., Thorpe, R., and Stebbings, R. (2010). Monoclonal antibody TGN1412 trial failure explained by species differences in CD28 expression on CD4+ effector memory T-cells. *British Journal of Pharmacology*, 161(3):512–526.

Endo, A. (2010). A historical perspective on the discovery of statins. *Proceedings of the Japan Academy, Series B*, 86(5):484–493.

Endo, A., Yoshio, T., Masao, K., and Kazuhiko, T. (1979). Effects of ML-236B on cholesterol metabolism in mice and rats: Lack of hypocholesterolemic activity in normal animals. *Biochimica et Biophysica Acta (BBA)—Lipids and Lipid Metabolism*, 575(2):266–276.

Fiorentino, A. R. and Dammann, O. (2015). Evidence, illness, and causation: An epidemiological perspective on the Russo-Williamson thesis. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 54:1–9.

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

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science*, 77:1070–1082.

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

Howick, J. (2011b). *The philosophy of evidence-based medicine*. Wiley-Blackwell, Chichester.

Howick, J., Glasziou, P., and Aronson, J. (2013a). Can understanding mechanisms solve the problem of extrapolating from study to target populations (the problem of 'external validity')? *Journal of the Royal Society of Medicine*, 106(3):81–86.

Howick, J., Glasziou, P., and Aronson, J. K. (2013b). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical Medicine and Bioethics*, 34(4):275–291.

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

Kenter, M. and Cohen, A. (2006). Establishing risk of human experimentation with drugs: lessons from TGN1412. *The Lancet*, 368(9544):1387–1391.

Levy, A. and Currie, A. (2015). Model organisms are not (theoretical) models. 66(2):327–348.

Longworth, F. (2006). Causation, pluralism and responsibility. *Philosophica*, 77:45–68.

Love, A. C. (2007). Functional homology and homology of function: Biological concepts and philosophical consequences. *Biology & Philosophy*, 22(5):691–708.

Maciejak, A., Leszczynska, A., Warchol, I., Gora, M., Kaminska, J., Plochocka, D., Wysocka-Kapcinska, M., Tulacz, D., Siedlecka, J., Swiezewska, E., Sojka, M., Danikiewicz, W., Odolczyk, N., Szkopinska, A., Sygitowicz, G., and Burzynska, B. (2013). The effects of statins on the mevalonic acid pathway in recombinant yeast strains expressing human HMG-CoA reductase. *BMC Biotechnology*, 13(1):68.

Parkkinen, V.-P. (2016). Robustness and evidence of mechanisms in early experimental atherosclerosis research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 60:44–55.

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

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

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

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

Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*.

Steel, D. (2008). *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford University Press, New York.

Steinberg, D. (2007). *The cholesterol wars: the skeptics vs. the preponderance of evidence*. Academic Press, San Diego, Calif, 1st edition.

Weber, E. (2007). Social mechanisms, causal inference, and the policy relevance of social science. *Philosophy of the Social Sciences*, 30(3):348–359.

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

Wimsatt, W. (2007). *Re-engineering philosophy for limited beings. Piecewise approximations to reality*. Harvard University Press, Cambridge, MA.

World Health Organization (2014). Ethical considerations for use of unregistered interventions for Ebola viral disease. Report of an advisory panel to WHO. Technical report, World Health Organization.