Synthese
https://doi.org/10.1007/s11229-018-1796-6

S.I.: EVIDENCE AMALGAMATION IN THE SCIENCES

Against external validity

Julian Reiss[1]

Received: 22 April 2017 / Accepted: 27 April 2018

Abstract Francesco Guala once wrote that 'The problem of extrapolation (or external validity as it is sometimes called) is a minor scandal in the philosophy of science'. This paper agrees with the statement, but for reasons different from Guala's. The scandal is not, or not any longer, that the problem has been ignored in the philosophy of science. The scandal is that framing the problem as one of external validity encourages poor evidential reasoning. The aim of this paper is to propose an alternative—an alternative which constitutes much better evidential reasoning about target systems of interest, and which makes do without (much) consideration of external validity.

Keywords External validity · Extrapolation · Methodology · Cancer causation · Evidence · Scientific reasoning

1 Introduction

Francesco Guala once wrote that 'The problem of extrapolation (or external validity as it is sometimes called) is a minor scandal in the philosophy of science' (Guala 2010: p. 1070). I agree with the statement, but for different reasons. The scandal is not, or not any longer, that the problem has been ignored in the philosophy of science. The scandal is that framing the problem as one of external validity encourages poor evidential reasoning. The aim of this paper is to propose an alternative—an alternative which constitutes much better evidential reasoning about target systems of interest, and which makes do without (or with a minimum of) considerations of external validity.

✉ Julian Reiss
julian.reiss@durham.ac.uk
http://www.jreiss.org

[1] Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK

Published online: 19 May 2018

In what follows, I will first describe the problem, sketch the main proposals for a solution we find in the literature today, note a common structure and argue that this way of thinking about the problem encourages poor evidential reasoning. More specifically, I will defend the thesis that thinking about causal inference in terms of the 'internal validity' and 'external validity' of causal claims encourages bad evidential reasoning because it suggests that for a claim to be externally valid of a target system of interest we have to establish an analogous claim for some experimental model system first—which, as I'll argue, is false. I will then propose my alternative according to which reasoning concerning target systems should begin with a *hypothesis about that system* and ask what types of evidence we shall need to establish that hypothesis. Finally, I'll go through a number of ways in which evidence from model systems can play a role in evidential reasoning without any consideration of whether or not some claim is 'externally valid'. The bulk of my examples will be drawn from the domain of cancer causation, and what I will say about an epistemology I call 'experimentalism' can mainly be found in large areas of biomedical research. But it has taken hold of pockets of the social sciences and economics too, and so my remarks find applicability in at least these domains within science.

2 What is the problem of external validity?

External validity is normally juxtaposed with internal validity, and the former defined in terms of the latter. Here is a short history of definitions (Shadish et al. 2002: p. 37; footnotes suppressed):

> Campbell (1957) first defined internal validity as the question, "did in fact the experimental stimulus make some significant difference in this specific instance?" (p. 297) and external validity as the question, "to what populations, settings, and variables can this effect be generalized?" (p. 297). Campbell and Stanley (1963) followed this lead closely. Internal validity referred to inferences about whether "the experimental treatments make a difference in this specific experimental instance" (Campbell and Stanley 1963, p. 5). External validity asked "to what populations, settings, treatment variables, and measurement variables can this effect be generalized" (Campbell and Stanley 1963, p. 5).

Others define the terms similarly, but the focus on causal claims is more explicit (Guala 2003: p. 1198; emphasis original):

> *Internal* validity is achieved when the structure and behavior of a laboratory system (its main causal factors, the ways they interact, and the phenomena they bring about) have been properly understood by the experimenter. For example: the result of an experiment E is internally valid if the experimenter attributes the production of an effect B to a factor (or set of factors) A, and A really is the (or a) cause of B in E. Furthermore, it is *externally* valid if A causes B not only in E, but also in a set of other circumstances of interest, F, G, H, etc.

In the biomedical sciences, the term 'extrapolation' is more common than 'external validity'. It describes the *inference* instead of a property of the inferred claim. Daniel

Steel gives a number of examples of such inferences and writes (Steel 2008: p. 3; emphasis original):

> In each of these cases, one begins with some knowledge of a causal relationship in one population, and endeavors to reliably draw a conclusion concerning that relationship in a distinct population. I will use the term *extrapolation* to refer to inferences of this sort.

The feature all these definitions have in common and on which I want to focus here is the basic understanding of the proper process of evidential reasoning. In this understanding, a first inference is made about one system—a laboratory or experimental system, or a test population. Then a second inference is made from knowledge about that system to knowledge or purported knowledge about another, related system or set of such systems—often a field system or a population different from the test population. Call the first kind a 'model system' and the second, 'target system(s) of interest'.

The 'problem of external validity' is the problem of making a reliable inference about target systems of interest when, for whatever reason, the target system isn't studied directly but indirectly by examining a model system or set of model systems. Hume's problem of induction can be understood as a version of this problem if by 'model system' we refer to some system in the present or past, say T, and the system of interest is the very same system at some point in the future, say $T+n$: Does the bread that has nourished me up until today continue to nourish me tomorrow? It is quite a different problem when the model and target systems differ in important respects: Is bread, which has proved to be nourishing to humans, safe for consumption for other species? Most discussions of external validity ignore the temporal aspect and instead focus on model-target inferences where model and target co-exist but differ in important respects.

What is an 'important respect'? To see that, first notice that the claims made about model and target system are normally causal claims, for instance about the toxicity of a substance, the effectiveness of a policy or the attribution of some outcome to the factor responsible for it. Causal claims in the social and biomedical sciences—in the context of which the problem of external validity is normally discussed—are true in virtue of some underlying structure or complex system or arrangement of mechanisms. For example, the fact that aflatoxin is toxic in humans is grounded in the human digestive system, and, generally, whether or not a compound is toxic depends on such a system. Likewise, the fact that increases in the money stock tend to bring about increases in prices has to do with certain socio-economic arrangements (concerning, for instance, a society's monetary constitution, banking system, and system of price determination), and, generally, whether or not increases in money cause increases in prices depend on such arrangements. A difference between systems is important if it pertains to one of its features that are responsible for the causal relation to hold. For instance, differences in digestive system would be important for the assessment of the external validity of toxicity claims; differences in monetary constitution for the assessment of the external validity of a claim about the power of money to raise prices.

The 'extrapolator's circle' (Steel 2008) obtains when there are important differences between model and target, and the target cannot be studied in the same way as the model. We often study model systems *because* it is possible to experiment on them

but not on the original target system. A model system might be more manipulable or an experiment on it cheaper to perform and there might be fewer ethical problems involved in model experiments. Importantly, however, it is rarely the case that the target system cannot be studied directly at all; rather, researchers *prefer* to go the indirect route because they expect epistemic benefits from doing so. However, this normally means that the model system differs in important respects from the target system (if it didn't, we would study the target directly). And this in turn means that the inference from model to target is not straightforward.

In sum, the terms external validity and extrapolation pertain to inferences in situations where the ultimate interest is in learning something about a specific target system of interest, where this system is, for one reason or another, epistemically less accessible than another related but importantly different system, and where we learn about the system of ultimate interest through learning about the related but importantly different system.

The catch is of course that the inferences of the second kind, i.e., the extrapolations, are far from trivial. If a substance is toxic for a non-human animal species, there is no guarantee that it is also toxic for most other species and, in particular, for humans. This is because different animal species have different digestive systems, and therefore there is variation (or heterogeneity) with respect to the toxicity of substances among species. If humans behave in a certain way in a laboratory situation, there is no guarantee that they will continue to behave in the same way in natural situations, for instance because of the Hawthorne effect (a behavioural modification of individuals in response to being observed; also called the 'observer effect'). And of course there is cultural variation in behaviour between different groups.

3 Solution strategies

In this section I will briefly survey the solution strategies that have been offered in the methodological literature. For each strategy, I will outline the essence of the strategy, state how it is supposed to solve the problem of external validity and finally describe its limitations.

3.1 Simple induction

This 'strategy' I mention only for the sake of completeness and because it appears to be used by some practitioners, but it is not so much a strategy as a denial that a strategy is needed. The proposal is: Maintain that a claim C_M, which has been established relative to model system M, also holds in target systems T_1, T_2, …, T_n, unless there are compelling reasons for not doing so (e.g., Post et al. 2013: pp. 641–642). So here the idea is essentially that the burden of argument is on those who raise the problem by asking them to provide reasons, compelling reasons even, to believe that a claim cannot be generalised (for a similar 'shift in the burden of proof' argument, see Starmer 1999: p. 15).

The proposal is either empty or leads to unreliable inferences if followed, depending on how strong the reasons for caution are assumed to be. If any old reason would do, the

proposal is empty as there are always differences between model and target systems, which can be held to constitute reasons for blocking the inference. Importantly, simple induction does not provide any information about what else to do in situations in which the simple induction on its own is likely to fail.

If 'reason' indeed means 'compelling reason', the proposal leads to many bad inferences. Sure, for some problems the relevant system features are so widely shared that one can confidently extrapolate from model to target. That parachute use is effective in preventing major trauma due to gravitational challenge has been established in numerous experiments (Smith and Pell 2003) and can safely be extrapolated to groups of patients who differ markedly in all sorts of respects. The effect is so large that it swamps whatever differences there are in the fragility of bones between humans. Similarly, countless experiments during the French Revolution have established the claim 'Decapitation causes death in humans'. Even if that population was highly selected, there are no reasons to believe that the claim cannot be generalised to other human, and indeed, other mammal and bird populations. Even Klaus Störtebeker, the German pirate about whom legend has it that, after he and his crew have been sentenced to death by beheading, he asked the mayor of Hamburg to release as many of his companions as he could walk past after being beheaded, the mayor granted the request and Störtebeker walked past 11 crew members, eventually died.

Generally, however, there will be some reasons but rarely compelling reasons for blocking the inference. If a compound is toxic in a non-human animal population, there are no compelling reasons to think that the claim cannot be extrapolated. Because the heterogeneity with respect to toxicity is considerable among species, we simply don't know whether one animal species or another is a good model for humans.

3.2 Analogy

One way to characterise extrapolation is as a species of analogical reasoning (LaFollette and Shanks 1997; Guala 2005). Following Paul Thagard (Thagard 1999), Guala argues that animal model-to-human inferences can be reconstructed as analogical inferences of the following kind (Guala 2005: p. 196):

(1) humans have symptoms Y,
(2) laboratory animals have symptoms Y,
(3) in laboratory animals, the symptoms are caused by factor X,
(4) the human disease is therefore also caused by X.

Thus far, the analogical approach does not make any advance relative to simple induction. Because of population heterogeneity, there is no reason to believe that a claim '*X* causes *Y* in animal model *M*' generally entitles researchers to make the inference to "*X* causes *Y* in the target population humans'. According to Guala, animal models have to be 'relevantly similar' to the human targets. Thus, extrapolation inferences (Guala 2005, p. 199; emphasis added):

> cannot be strong unless experimental and field evidence have been generated by systems that are similar in *all relevant* respects or, in other words, unless *all*

> sources of external validity [i.e., extrapolation] error have been taken care of by means of accurate design.

The problem is that this strict demand will lead to a sceptical conclusion according to which extrapolations are never justified because animal models are never similar in *all* relevant respects, which is why we experiment on animals to begin with (for the sceptical conclusion, see LaFollette and Shanks 1997; for a discussion, Steel 2008: pp. 92–99).

In more recent work, Guala demands merely that model and target be similar in respects 'that are deemed relevant by our current background knowledge' (Guala 2010: p. 1075). Background knowledge tells us that certain variables may be correlated with the effectiveness of a cause to bring about an effect or with the likelihood that a causal relationship holds. A drug may work for men but not for women, and it may decrease in efficacy with age. The proposal then is to measure known co-variates in a model or (better) a set of models, determine the value of each co-variate in the target, and to form an expectation about the causal claim on the basis of the model estimates.

The method is similar to the measurement of 'hedonic' price indices in econometrics. (Hedonic regression is used more widely in economics and other social sciences, so this serves just as a comparison with a well-known example.) A new version of some good g is introduced in the market at time t and a price p_{gt}. This price differs from the price of the good in the last period p_{gt-1}. The pure price change (p_{gt}/p_{gt-1})–1 is likely to overestimate inflation because of the change in its quality. That, however, is not directly observable. Background knowledge tells us what factors (x) are likely to be relevant. If, say, g is an operating system, these factors may be speed, reliability, number of features, user friendliness etc. The contributions of each of these to the price can now be determined by estimating a factor weight from data on existing goods via regression $p_{gj} = \alpha + \sum_i \beta_{ji} x_{ij}$, where the g_j's are the existing goods and the x_i's the quality-determining factors. A nice feature of the method is that it is quite flexible. Suppose that user friendliness is the factor consumers really care about, but it's not measurable or not measured. 'Lines of code' is something consumers don't care about at all, but it is, for whatever reason, correlated with user friendliness. In this case, lines of code can feature in the regression as one of the x_i's even though it is not a 'cause' of quality. The price the new good should have on the basis of its changed quality can now be estimated and thus, in turn, the pure price change.

There is an ontological and an epistemic prerequisite for the method to work. It must in fact be the case that the causal effect of one variable on another can be represented by a regression equation like the above. It doesn't have to be a simple linear average of the factor contributions, but there must be some principle of combination, even if it is highly case specific. There are some reasons to believe that there is no general principle of combination, i.e., that factors sometimes interact so that the relationship between a given cause and its effect depends on what else is present in the situation. That, say, a specific substance is toxic can be explained with reference to the organism's digestive system. But why should toxicity across species be related to features of organisms in a regular manner? Being subject to common laws of physics and chemistry may explain some commonality, but in general we would expect organisms to solve problems of adaptation to their environments within their ecological niche.

Across species generalisations are therefore unlikely to be very reliable. In addition we must have enough data to estimate the factor weights reliably. The more factors there are, the more experiments have to be run. In practice, the second prerequisite will be particularly constraining because of the high cost of running experiments.

3.3 Comparative process tracing

As mentioned above, causal relations in the biomedical and social sciences hold on account of an underlying structure or complex system. Structures or complex systems can in turn be understood as arrangement of mechanisms. Process tracing comprises techniques for learning such arrangements of mechanisms (Darden and Craver 2002; Glennan 2005). Daniel Steel has developed an account of extrapolation that uses process tracing for model-target inferences (Steel 2008).

Because model systems are epistemically accessible, researchers are often able to establish not only that some variable causes another in the model system, but also how, i.e., through what mechanism or mechanisms. Comparative process tracing proceeds on the principle, 'Same causal process—same effect'. Thus, If C causes E in M through some mechanism Q, and we find Q also operative in the target T, we infer that C causes E in T as well.

However, since T is not as accessible, it will be difficult to establish the full mechanism in T—which is why the experiment is conducted on M in the first place. Steel now argues that the extrapolation can proceed on a relatively rudimentary understanding of the mechanism in T. Researchers just have to know at what point or points the mechanisms in M and T are likely to differ. When differences are suspected, these differences will be transmitted downstream from the likely point of departure. Hence it is only necessary to examine the mechanism Q at points downstream from where differences between model and target are suspected.

The main problem with this proposal is that the epistemic requirements are really high. Mechanisms, especially in biology and social systems, tend to be so complicated that they are hard to learn even in models. In biology, our ability to learn mechanisms has greatly improved with the advancement of molecular biology, but it remains the case that the underlying mechanisms for accepted causal relations are not understood, or that they come to be understood only much after the macro causal claim has been accepted (Reiss 2012). In the social sciences, gaining knowledge of mechanisms that is robust enough as to be useful for extrapolation is severely compromised—and Steel admits as much (Steel 2008: p. 173; see also Reiss 2010).

3.4 External validity by engineering

This method too is based on the principle 'Same causes—same effects' but it works not by simply comparing the underlying structure for a causal relation between model and target but instead by engineering the target. Guala argues that this approach was used in case of the Federal Communications Commission (FCC) auctions of the electromagnetic spectrum (Guala 2005: Ch. 8). Laboratory experiments concerning the effects of different auction rules had to be conducted because the rules were too com-

plex for theoreticians to make definite predictions. Different rules were thus tested in the laboratory, and the real auctions followed the rules that worked best in the lab.

The engineering strategy, in short, is this: build your institution in such a way as to mimic the experimental conditions as closely as possible. It is not the case that existing systems are analysed and explained but rather new systems are created, following the recommendations derived from theory and experiments. The approach works, at best, only when the design of new institutions is the aim (and not when existing systems are sought to be explained). Further, the approach works only to the extent that an institution that closely mirrors the experiment can be created. This seems to have been the case in the FCC auctions, but that this is possible is not guaranteed. Finally, important differences between lab and the field will remain and they may matter.

3.5 Field experiments

Field experiments are common in economics, where problems of external validity are particularly pressing because of the high degree of malleability of human behaviour, which makes it difficult to extrapolate straightforwardly from the lab to the wild. Field experiments are said to build a bridge between naturally occurring systems and laboratories (List 2007) because they employ individuals in their usual environs, albeit with some degree of control and, usually, two or more groups that receive different treatments for comparison.

Laboratory experiments tend to produce results that have a good chance of being internally valid because the degree of control researchers can exercise is very high. However, the downside of the ability to exercise control is the creation of experimental artefacts such as the Hawthorne effect. Observational studies are minimally invasive and thus unlikely to create experimental artefacts, but they are always subject to possible biases such as selection bias (where individuals self-select into 'treatment groups' so that it is always possible that their reason for selecting a treatment is correlated with the outcome). Where lab experiments and observational studies constitute the opposite end poles of the spectrum, different kinds of field experiments lie in between (see Harrison and List 2004 for a taxonomy).

The bridge metaphor is apt in this sense, but it would be mistaken to assume that Aristotelian moderation solves inferential problems. A little bit of experimental artefact and a little bit of selection bias might in fact be worse than anything created by the extremes. In particular, there is no guarantee that a series of experiments of increasing realisticness converges to the result that is true of the target population. For this to happen, interactive effects between the true effect and the various study biases would have to be absent, but there is no guarantee that this is so. If anything, there is considerable evidence that factors in the social world, especially behavioural factors, tend to be interactive rather than additive (Reiss 2008, 2017).

4 Thinking in terms of external validity encourages bad evidential reasoning

That all the strategies for extrapolation offered in the literature come with conditions for application and limitations is not a reason not to use them where they work and with caution. The problem, in my view, lies at a deeper level. Understanding inferences about a target system of interest as a problem of extrapolation or securing external validity encourages foundationalist thinking about inference. All methods discussed in the previous section ultimately aim to learn about a target system T and proceed by examining a different but related system M and inferring about T from what has been learned about M. The reason lies hidden in the alleged 'epistemically benefit' mentioned near the end of Sect. 2. As discussed there, it is not that target systems cannot be studied directly. However, for a variety of reasons—ethical, financial, technological—they cannot, or not fruitfully, be experimented on. It is the unavailability of experimentation that calls for a detour via a surrogate system.

Let's call 'experimentalism' the view that randomised experiments are the gold standard of causal inference. This view is widely held in the biomedical sciences (going back to biostatisticians such as Ronald Fisher, Joseph Berkson, Jacob Yerushalmy and others; see Parascandola 2004) and, more recently, also across the social sciences (Shadish et al. 2002; for economics, see for instance Angrist and Pischke 2010). We can distinguish between a conservative and a liberal form of experimentalism. Conservative experimentalism regards only the gold standard as intrinsically reliable and dismisses all other methods as unreliable. Liberal experimentalism defines a hierarchy of methods of causal inference. If randomised experiments are the gold standard, then other methods can be ranked with respect to reliability according to how closely they mimic randomised experiments. Liberal experimentalists thus recommend the use of 'quasi-experimental' techniques should the gold standard be unavailable.

I am not asserting here that one has to be an experimentalist in this sense in order to understand reasoning about target systems as a problem of ascertaining external validity. However, it is difficult to see the motivation behind this way of understanding reasoning about target systems without an experimentalist framework. If one is ultimately interested in learning the truth of a hypothesis of the form 'C causes E in T' (the 'T-hypothesis'), why would one make a detour via learning the hypothesis 'C causes E in M' first (the 'M-hypothesis')? The only sensible answer appears to be: because we can learn the M-hypothesis more reliably.

To see this, suppose that each scientific method (randomised trials, non-randomised intervention studies, prospective cohort studies, retrospective case-controlled studies, etc.) confers a fixed, context-independent probability on the hypothesis tested using that method. Thus, after testing using method m_i, the T-hypothesis, h^T, has the probability $P(h^T{:}m_i) = p$, where $P(h{:}m)$ refers to the probability of hypothesis h given it has been tested by method m. The same is true of the M-hypothesis: $P(h^M{:}m_j) = q$. i may or may not be the same as j.

The inference from the M-hypothesis to the T-hypothesis is an inference by analogy (Guala 2010), thus an inductive inference, and thus itself uncertain. Let us suppose too that these inferences also come with a fixed probability: $P(h^M \rightarrow h^T) = r$. Now, if the goal is to maximise the probability of the T-hypothesis to be true, we will (a)

choose the method that maximises p (or both p and q if we do extrapolate) and (b) use extrapolation if and only if $rq > p$.

For the defender of extrapolation, this is a tall order. First, given that no method or extrapolation is perfectly reliable (i.e., $p, q, r < 1$), the extrapolation must on its own be more reliable than the best method with which we can study the target directly. Second, depending on the exact number r, this requires an potentially highly reliable method to study the model q. To give a numerical example that is very charitable to the defender of extrapolation, suppose $r = .85$. Then, if the best method with which we can study the target directly confers a probability of .7, the reliability of the method with which the model is studied must be about .82. Essentially, because the extrapolation compounds two sources of uncertainty—the inference concerning what is true of the model and the inference from model to target—both components had better be individually rather reliable.

Thus, the conservative experimentalist will recommend extrapolation whenever the target cannot be studied experimentally for ethical, practical or financial reasons; the liberal experimentalist when the combined reliability rq of extrapolating the experimental result from the model exceeds the reliability of the best applicable quasi-experimental method p.

There is also historical evidence of the connection between external validity reasoning and experimentalism. The term 'external validity' originated within an experimentalist framework. As we've seen above, the distinction between internal and external validity is due to Campbell (1957). Donald Campbell was clearly an experimentalist in the sense used here, and the Campbell Collaboration, a non-profit that aims to improve decision-making about the effects of interventions in the social, behavioural, and educational arenas and was named after Donald Campbell, reviews social science evidence on the basis of experimentalist principles.

Experimentalism is what I call a foundationalist methodology. A foundationalist methodology maintains that certain methods are *intrinsically* reliable; the use of these methods does not stand in need of epistemic justification by considerations related to the application of the method to a particular case. Beliefs tested by an intrinsically reliable method are justified. Call these 'basic beliefs'. Other beliefs must, in order to be justified, be derived from the basic beliefs. There are, then, two primary modes of justification according to a foundationalist methodology: a basic belief is justified by its having passed a test by the intrinsically reliable method; a non-basic belief is justified by being derived from a basic belief. According to experimentalism, the anchoring of non-basic beliefs in basic beliefs happens through extrapolations.

What is an 'intrinsically reliable' method? As we have seen, conservative and liberal experimentalists differ in their answer. The conservative experimentalist regards *only* randomised experiments as intrinsically reliable. This view has some plausibility. According to a recently popular theory of causation, causation *is* invariance under (ideal) intervention (Woodward 2003). An ideal intervention (on C with respect to E) has the following properties (Woodward 2003: p. 98; notation slightly altered for consistency):

I1 *I* causes *C*.
I2 *I* acts as a switch for all the other variables that cause *C*. That is, certain values of *I* are such that when *I* attains those values, *C* ceases to depend on the values of other variables that cause *C* and instead depends only on the value taken by *I*.
I3 Any directed path from *I* to *E* goes through *C*. That is, *I* does not directly cause *E* and is not a cause of any causes of *E* that are distinct from *C* except, of course, for those causes of *E*, if any, that are built into the *I*-*C*-*E* connection itself; that is, except for (a) any causes of *E* that are effects of *C* (i.e., variables that are causally between *C* and *E*) and (b) any causes of *E* that are between *I* and *C* and have no effect on *E* independently of *C*.
I4 *I* is (statistically) independent of any variable *Z* that causes *E* and that is on a directed path that does not go through *C*.

An ideal randomised trial implements an ideal intervention. That can easily be verified:

I1 The randomisation device (or random number) causes the treatment status.
I2 Clinical control makes sure that patients receive the treatment only from the trial and not from elsewhere.
I3 Patients, experimenters, and possibly others are blinded with respect to treatment status in order to make sure that the allocation to treatment group affects the outcome only through the treatment itself.
I4 Since the allocation to treatment group is done by a random device, allocation should be statistically independent of other causes of treatment.

Given that a(n ideal) randomised experiment implements an ideal intervention, it is no surprise that the (ideal) randomised experiment is a reliable means to test for causation if what we mean by causation is invariance under ideal intervention. But of course, and here comes the snag, the interventionist theory is only one theory among many, though popular among contemporary philosophers certainly not universally or even widely accepted, and there are in fact good reasons to reject it (see Reiss 2015a: Chs 1 and 9 for some such reasons). Randomised experiments are not generally intrinsically reliable methods to test for causation under alternative theories of causation. It is not clear, to say the least, whether randomised experiments come out as intrinsically reliable under a process account, for instance—because randomised experiments do not test for the existence of a process from *C* to *E* (and thus a positive test result does not guarantee the existence of such a process).

The liberal experimentalist faces a second problem. She has to show in addition that 'quasi-experimental' methods are reliable. And this is a difficult endeavour, even under an interventionist interpretation of causation. To give an example, the estimation of instrumental variables is such a quasi-experimental method. And there is indeed some sense in which a valid instrument is 'similar' to Woodward's ideal intervention (for a precise account of similarities and differences, see Reiss 2005). But everything is similar to everything else in countless ways, and the experimentalist does not provide an account of what differences matter. Instruments, for instance, do not act as 'switches', i.e., a typical instrument violates Woodward's condition I2. Does that invalidate instruments as method of inference? If not, why not, and how much of the

other conditions can be violated for the method to no longer count as reliable? The experimentalist has no answers to these questions.

The main problem for both conservative and liberal experimentalist is, however, that no *real* randomised experiment implements an *ideal* intervention. Here is a selection of reasons why:

- for finite test populations, treatment and control group(s) may always be unbalanced with respect to some prognostic factors, especially when the number of such factors is high (in violation of I4);
- blinding is not possible for all types of treatment, and even when possible, treatment status can be revealed by its effectiveness (in violation of I3);
- treatments can be 'leaked', for instance, when different groups share treatments (in violation of I2);
- attrition rates can be different between treatment groups and correlated with prognostic factors (also in violation of I4).

That *real* randomised experiments are an intrinsically reliable method is thus untenable. When it comes to assessing a result, a case has to be made that:

- treatment and control group(s) were indeed (likely to be) balanced;
- all participants were successfully blinded with respect to treatment status or else that lack of blinding did not confound the result;
- treatments have not been leaked;
- attrition rates were similar between groups or, if dissimilar, uncorrelated with prognostic factors;
- and many other sources of error that haven't been mentioned here have been controlled.

If beliefs in the results of randomised experiments are reliable, they are reliable relative to a backdrop of other beliefs about the situation at hand and not because of what they are. Hence, foundationalism's answer to the first question concerning the justification of the basic beliefs is unsatisfactory. The answer to the second question concerning the derivation of non-basic beliefs from the basic beliefs has already been given in Sect. 3. Even if, against what I just argued, experimentalism's basic beliefs were justified, inferences to other beliefs are very unlikely to be reliable because extrapolation is so poorly understood. Methodological foundationalism thus gives an unsatisfactory answer also to the second question.

Methodological foundationalism is a flawed epistemology. I should perhaps say explicitly here that I do not accuse any philosopher of science of subscribing to a flawed epistemology. Methodological foundationalism is, however, widely held across the biomedical and social sciences, and by trying to rationalise certain inferential strategies employed by scientists in my view some philosophers make themselves complicit in poor reasoning about target systems. It may well be true that work by philosophers on the 'problem of external validity' would never have existed if scientists did not employ this mode of reasoning. However, that doesn't change the fact that the reasoning is poor indeed and only really makes sense from the point of view of a flawed epistemology. Philosophers could have pointed that out.

Let me therefore propose an alternative. Methodological foundationalism asks, 'What can we know with a high degree of certainty, and how do we learn what we

want to know from what we can know with a high degree of certainty?' instead of asking what we need to know in order to make a reliable causal inference about a target system of interest. It is sometimes argued that this way of thinking privileges certainty over relevance (Cartwright 2009). Milton Friedman once said, 'A society that puts equality before freedom will get neither. A society that puts freedom before equality will get a high degree of both'. Analogously, I maintain that an epistemic community that puts certainty before relevance will get neither (for the reasons discussed). An epistemic community that puts relevance before certainty will get a high degree of both. In what follows I will sketch an alternative, pragmatist epistemology and demonstrate that the second part of the slogan is correct.

5 Non-foundationalist reasoning about target systems of interest

One major alternative to foundationalism is contextualism. At the most general level, contextualism maintains that justification is relative to context. In previous work I have defended that the relevant context is an *epistemic* context given by substantive information about the case at hand, the nature and purpose of the inquiry, and relevant methodological, conceptual, and ethical norms (Reiss 2015b).

A causal inquiry begins with a causal hypothesis about a target system of interest: 'Vinyl chloride causes cancer of the liver in humans', 'HRT causes breast cancer in women who have had a hysterectomy', 'Ultraviolet radiation causes melanoma in people of colour'. A fundamental distinction the framework makes is between the *support* for a hypothesis and its *warrant*, a distinction most alternative accounts conflate in the notion 'evidence'. Support comprises any information that is relevant to assessing the hypothesis. It is synonymous with the notion of a *piece of evidence* (such as a fingerprint on the murder weapon). Warrant, by contrast, refers to the degree of justification of the hypothesis, which is based on the entire *body of evidence* (which comprises *all* the information used in the judgement). Bayesianism, to give just one example, conflates the two notions because it defines as support anything that changes the degree of warrant: e is evidence (in the sense of support) for H if and only if $P(H \mid e) > P(H)$, i.e., the degree of warrant of H is higher after e has been learned. (We can't use the likelihood $P(e \mid H)$ to define support as distinct from warrant because $P(e \mid H) > P(e)$ if and only if $P(H \mid e) > P(H)$.) Collecting information relevant to assessing a hypothesis and making up one's mind about it are two different processes, however, which should be kept separate.

Support for a hypothesis falls into two kinds: direct and indirect. Direct support addresses the question: 'What patterns in the data are researchers entitled to expect under the supposition of the truth of the hypothesis?' Examples of such patterns include correlations between the cause and effect variables in the population of interest or causal process observations that provide evidence for a mechanism from cause to effect. Examples of causal process observations in cancer causation are observations of DNA damage, gene mutation, sister chromatid exchange, micronucleus formation, chromosomal aberrations and aneuploidy. Indirect support addresses the question: 'What patterns in the data incompatible with alternative accounts of the direct evidence are researchers entitled to expect?' A correlation between cause and effect

variable, for instance, can be accounted for by a causal relation going from the putative effect variable to the cause, a common cause, or numerous biases such as selection bias, experimenters bias, diagnostic error, the Berkson paradox and so on. Any pattern in the data incompatible with these alternative hypotheses constitutes a piece of indirect support. An example is given by a study that showed that under the alternative hypothesis 'diagnostic error' the error in diagnoses of death from lung cancer would have to be an order of magnitude greater among men than among women, and among older individuals than among younger individuals (Gilliam 1955). These patterns were, together with the background information that such large and systematic differences in diagnostic error are extremely unlikely, taken to rule out the diagnostic error hypothesis and thus a piece of indirect support.

Within this pragmatist framework, justification is contextual in that it derives from contextual factors such as the existing domain-specific information an epistemic community has about the situation at hand (for example about what patterns in the data causal relations typically produce in that domain, what the relevant alternative hypotheses are, about the operation of institutions such as hospitals and the behaviour of individuals such as patients, doctors and pathologists), the nature and purpose of the inquiry (an inquiry would look for different evidence depending on whether the goal is, say, explanation or public health) and relevant norms (for example, standards will vary with stakes).

Unlike foundationalism, contextualism takes no type of information, method or study as intrinsically reliable or trustworthy. If there are no reasons to dispute the results of a study published in a respectable medical journal, researchers are entitled to take them as given. However, that can change when information about, say, irregularities in the peer reviewing process or fraudulent or sloppy behaviour of an individual researcher become available. To give another example: if we learn that 'Most Published [Biomedical] Research Findings are False' (Ioannidis 2005), the ability to draw on existing work in any area of biomedical research is obviously compromised and will be necessary to examine any given study for frequently occurring errors such as low power, design problems, a host of biases, conflicts of interest and so on.

The inquiry ends with a conversion of the direct and indirect support into a judgement about the degree to which the hypothesis is warranted. In Reiss (2015a, b) I distinguish four degrees (in decreasing order of strength): proof, strong warrant, moderate warrant, and weak warrant. Degree of warrant is, essentially, proportional to the number and significance of the relevant alternative hypotheses that have been ruled out.

Randomised experiments play no special role in the pragmatist theory. Randomisation is one way to control for selection bias (Worrall 2002). Selection bias is an alternative hypothesis that can account for a correlation, and controlling for it means to raise the degree of warrant for the causal hypothesis. But there are alternative ways. If background information tells us that individuals who are prone to contracting lung cancer are likely to select into the population of smokers, then a correlation between smoking and lung cancer cannot provide strong warrant for the causal hypothesis. But if background information also tells us that (a) proneness to lung cancer is most likely to have a genetic basis; (b) genetic factors explain about 20% of cancer risk;

(c) (strong) smokers have 6000-fold increased risk as compared to non-smokers, then selection bias can be ruled out without randomisation.

6 Contextualist reasoning from models without external validity

Within the pragmatist theory sketched above information drawn from models, i.e., from systems that resemble the target system of interest in relevant respects and stand in for them, can play a variety of roles. Let me focus on a number of heuristic and evidential roles here.

6.1 Suggesting hypotheses

Most hypotheses concerning the carcinogenicity of substances stem from a suspicion, based on clinical experience, that the concurrence of exposure to a particular agent and occurrence of a cancer has happened rather more frequently than would be expected by chance. The initial—direct—support is often a series of case reports, say, concerning factory workers exposed to the substance.

However, for some agents (e.g., diethylstilbestrol, melphalan, mustard gas, and vinyl chloride) evidence of carcinogenicity in experimental animals preceded any evidence obtained from studies on humans (Vainio et al. 1995). In such cases, information drawn from animal models can play the heuristic role of suggesting a hypothesis concerning humans. Importantly, in this role, information drawn from animal models does not provide any support or warrant concerning carcinogenicity in humans.

6.2 Providing direct support

Once the hypothesis is being investigated, the very same laboratory experiments on animal models can provide direct support. In order for the experiments to play this role, we need a bridge principle that connects carcinogenicity in animals with carcinogenicity in humans. The bridge principle is the following (IARC 2006: p. 12; emphasis added):

> All known human carcinogens that have been studied adequately for carcinogenicity in experimental animals have produced positive results in one or more animal species.

Due to this piece of background information, a positive animal study provides direct support for carcinogenicity in humans. This is because, given the information, under the supposition of the truth of the human carcinogenicity hypothesis researchers are expected to find positive results in one or more animal species. Considerations of external validity play absolutely no role here. It does not matter whether the model is a 'good' one, whether salient mechanisms are shared or whether the principle 'same causes—same effects' can be applied. Animal studies are direct support because if the substance is carcinogenic in humans, it will be carcinogenic in some animal species. Whereas in 6.1 the animal study must precede the inquiry of carcinogenicity in humans,

in 6.2 the order of events is irrelevant. It is also important to note that direct support alone does not provide any warrant for the hypothesis, not even what I call 'weak warrant', for which at least some alternative hypotheses have to be ruled out.

This role can be significant when other types of direct support are unavailable. Cause variables are not always correlated with their effect variables, for instance. If so, the inquiry has to proceed from other kinds of direct support and, for human carcinogenicity studies, experiments on animal models can fit the bill.

6.3 Specifying hypotheses

Hypotheses of the form 'Substance *S* causes cancer in humans' are very unspecific because cancers can occur at numerous sites and, at each site, have a variety of morphologies. It is often difficult to rule out alternative hypotheses concerning the causes of cancer because cancer can have a very large number of causes. Liver cancer, for example, is caused by birth defects, alcohol abuse, aflatoxins, chronic infection with liver diseases such as hepatitis B and C, hemochromatosis, cirrhosis, obesity and diabetes, fatty liver disease, and other factors. It will be difficult to find data sets that control for all these factors, experiments testing for toxicity are unethical and even if one could conduct them, they would be unlikely to provide reliable information due to the large number of plausible confounders.

Information about cancer sites and morphologies from animal studies can help to form more specific expectations about humans. Angiosarcomas of the liver, for instance, have been observed in mice and rats exposed to vinyl chloride in lab experiments (IARC 1974). Subsequently, the same condition was found in humans exposed to the substance. Angiosarcomas are extremely rare, however, so that an observed correlation between exposure to vinyl chloride and incidents of the disease can hardly be accounted for by other hypotheses.

Again, the hypothesis was not tested for external validity. That is, the evidence on the model is not used to justify the hypothesis about the target. Rather, the information from animal experiments helped to formulate a more specific hypothesis, and alternatives could be ruled out on the basis of human data. This is a general feature of causal reasoning: the more specific researchers' expectations are concerning effect size, the temporal evolution of the effect, its modus operandi, and its form, the easier it is to rule out alternatives hypotheses. An alternative might be able to account for a qualitative effect but not for a specific size or for the occurrence of the effect but not for when and where it occurred.

6.4 Analogies with known carcinogens

Analogies with known carcinogens support inferences that are somewhat similar to extrapolations, but they have a different form. Extrapolations have the form '*C* causes *E* in *M*, therefore *C* (also) causes *E* in *T*', where *T* is relevantly similar. Analogies have the form '*C* causes *E* in *T*, therefore *C** (also) causes *E* in *T*'. They can play an inferential role provided the modus operandi of *C* to cause *E* in *T* is well understood.

Polybrominated biphenyls (PBBs) are structurally very similar to the known carcinogens of the polychlorinated biphenyl (PCB) group. Because of this, and because it is known that structurally similar compounds act through the same mechanisms of carcinogenesis, if PBBs were carcinogenic, we would expect them to operate through the same mechanisms as PCBs (IARC 2016: p. 37):

> Concerning experimental and mechanistic studies, while there is an extensive body of literature on the carcinogenicity of PCBs, their brominated analogues have received much less attention and study. PBBs will likely be found to exhibit their toxicity and disease potential via many of the same pathways as their chlorinated counterparts, with equivalent or greater toxicity.

Thus, analogies provide direct support for carcinogenicity. Again, no extrapolation is needed here, but an examination of the mechanisms through which PBBs operate in humans.

7 Conclusions

This paper argued in favour of two claims. First, thinking about causal inference in terms of the 'internal validity' and 'external validity' of causal claims encourages bad evidential reasoning because it suggests that for a claim to be externally valid of a target system of interest we have to establish an analogous claim for some experimental model system first. That is not so. Reasoning concerning target systems should begin with a hypothesis about that system and ask what types of evidence we need to establish that hypothesis. The pragmatist theory of evidence sketched in Sect. 5 provides an account that proceeds in that way. Second, within such an account of reasoning about target systems, information drawn from model systems can play a variety of heuristic and evidential roles that have nothing to do with extrapolations or judgements of external validity.

This paper has focused on heuristic roles and the provision of direct support. I am not suggesting that evidence that draws on models such as animal models can never play a justificatory role, i.e., also provide warrant for a hypothesis. Indeed, in the IARC classification system, animal evidence can help to justify a hypothesis concerning humans:

> **Group 2A: The agent is probably carcinogenic to humans**
> This category is used when there is limited evidence of carcinogenicity in humans and sufficient evidence of carcinogenicity in experimental animals. In some cases, an agent may be classified in this category when there is inadequate evidence of carcinogenicity in humans and sufficient evidence of carcinogenicity in experimental animals and strong evidence that the carcinogenesis is mediated by a mechanism that also operates in humans.

> **Group 2B: The agent is possibly carcinogenic to humans**
> This category is used for agents for which there is limited evidence of carcinogenicity in humans and less than sufficient evidence of carcinogenicity in experimental animals.

Thus, if there is 'limited' evidence on humans, a compound can be bumped up from 2B to 2A if there is 'sufficient' evidence on animals. Surveying a large number of IARC monographs indicates, however, that such upgrades are very rare, in particular in older monographs. An explanation for this is that stronger bridge principles than those of the kind discussed in Sect. 6 are needed. We would need to know not just 'if *S* is carcinogenic in humans, then it will be carcinogenic in some animal species' but something much stronger. It is well possible that when better bridge principles are available, animal evidence will gain in significance in warranting hypotheses concerning humans and that some of the epistemic strategies discussed in Sect. 3 can be recast within the pragmatist framework. For now, I will leave this to a future paper.

Acknowledgements I wish to thank audiences at the University of Kent and Copenhagen as well as three anonymous referees for most valuable comments on earlier drafts of this paper. Funding from the Spanish Ministry of Science and Innovation for the research project 'Laws, explanation, and realism in physical and biomedical sciences' (FFI2016-76799-P) and the European Research Council for the project 'Knowledge for Use' (Grant agreement No. 667526 K4U) is gratefully acknowledged.

References

Angrist, J., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives, 24*(2), 3–30.

Campbell, D. (1957). Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin*, *54*, 297–312.

Campbell, D., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Cartwright, N. (2009). Evidence-based policy: What's to be done about relevance. *Philosophical Studies, 143*(1), 127–136.

Darden, L., & Craver, C. (2002). Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences, 33*(1), 1–28.

Gilliam, A. (1955). Trends of mortality attributed to carcinoma of the lung: Possible effects of faulty certification of deaths due to other respiratory diseases. *Cancer, 8,* 1130–1136.

Glennan, S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 443–464.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science, 70,* 1195–1205.

Guala, F. (2005). *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science, 77*(5), 1070–1082.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009–1055.

IARC. (1974). *IARC Monographs on the evaluation of carcinogenic risk of chemicals to man*. Volume 7: Some anti-thyroid and related substances, nitrofurans and industrial chemicals. Lyon: International Agency for Research on Cancer.

IARC. (2006). *IARC monographs on the evaluation of carcinogenic risks to humans: Preamble*. Lyon: International Agency for Research on Cancer.

IARC. (2016). *IARC monographs on the evaluation of carcinogenic risks to humans*. Volume 107: Polybrominated biphenyls and polychlorinated biphenyls. Lyon: International Agency for Research on Cancer.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

LaFollette, H., & Shanks, N. (1997). *Brute science: Dilemmas of animal experimentation*. London: Routledge.

List, J. A. (2007). Field experiments: A bridge between lab and naturally occurring data. *The B.E. Journal of Economic Analysis & Policy, 5*(2), 1–47.

Parascandola, M. (2004). Two approaches to etiology: The debate over smoking and lung cancer in the 1950s. *Endeavour, 28*(2), 81–86.

Post, P., de Beer, H., & Guyatt, G. (2013). How to generalize efficacy results of randomized trials: Recommendations based on a systematic review of possible approaches. *Journal of Evaluation in Clinical Practice, 19*(4), 638–643.

Reiss, J. (2005). Causal instrumental variables and interventions. *Philosophy of Science, 72*(PSA 2004), 964–976.

Reiss, J. (2008). *Error in economics: Towards a more evidence-based methodology*. London: Routledge.

Reiss, J. (2010). Across the boundaries: Extrapolation in biology and social science, Daniel P. Steel. Oxford University Press, 2007. xi + 241 pages. *Economics and Philosophy, 26*(3), 382–390.

Reiss, J. (2012). Third Time's a Charm: Wittgensteinian Pluralisms and Causation. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 907–927). Oxford: Oxford University Press.

Reiss, J. (2015a). *Causation, evidence, and inference*. New York, NY: Routledge.

Reiss, J. (2015b). A pragmatist theory of evidence. *Philosophy of Science, 82*(3), 341–362.

Reiss, J. (2017). Are there social scientific laws? In L. McIntyre & A. Rosenberg (Eds.), *The Routledge companion to philosophy of social science* (pp. 295–309). New York, NY: Routledge.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Smith, G., & Pell, J. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *British Medical Journal, 327,* 1459–1461.

Starmer, C. (1999). Experiments in economics: Should we trust the dismal scientists in white coats? *Journal of Economic Methodology, 6*(1), 1–30.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

Thagard, P. (1999). *How scientists explain disease*. Princeton, NJ: Princeton University Press.

Vainio, H., Wilbourn, J. D., Sasco, A. J., Partensky, C., Gaudin, N., Heseltine, E., et al. (1995). Identification of human carcinogenic risk in IARC Monographs. *Bulletin du Cancer, 82,* 339–348. **(In French)**.

Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.

Worrall, J. (2002). What evidence in evidence-based medicine. *Philosophy of Science, 69,* S316–S330.