Routledge
Taylor & Francis Group

Journal of Economic Methodology

ISSN: 1350-178X (Print) 1469-9427 (Online) Journal homepage: http://www.tandfonline.com/loi/rjec20

Extrapolation of causal effects – hopes, assumptions, and the extrapolator's circle

Donal Khosrowi

To cite this article: Donal Khosrowi (2019): Extrapolation of causal effects – hopes, assumptions, and the extrapolator's circle, Journal of Economic Methodology

To link to this article: https://doi.org/10.1080/1350178X.2018.1561078

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 07 Jan 2019.

Submit your article to this journal

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=rjec20

JOURNAL OF ECONOMIC METHODOLOGY
https://doi.org/10.1080/1350178X.2018.1561078

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Extrapolation of causal effects – hopes, assumptions, and the extrapolator's circle

Donal Khosrowi

Department of Philosophy, Centre for Humanities Engaging Science and Society, Durham University, Durham, UK

ABSTRACT

I consider recent strategies proposed by econometricians for extrapolating causal effects from experimental to target populations. I argue that these strategies fall prey to the extrapolator's circle: they require so much knowledge about the target population that the causal effects to be extrapolated can be identified from information about the target alone. I then consider comparative process tracing (CPT) as a potential remedy. Although specifically designed to evade the extrapolator's circle, I argue that CPT is unlikely to facilitate extrapolation in typical econometrics and evidence-based policy applications. To argue this, I offer a distinction between two kinds of extrapolation, *attributive* and *predictive*, the latter being prevalent in econometrics and evidence-based policy. I argue that CPT is not helpful for predictive extrapolation when using the kinds of evidence that econometricians and evidence-based policy researchers prefer. I suggest that econometricians may need to consider qualitative evidence to overcome this problem.

ARTICLE HISTORY
Received 15 December 2017
Accepted 19 July 2018

KEYWORDS
Evidence-based policy; econometrics; extrapolation; extrapolator's circle; comparative process tracing; qualitative evidence; mechanisms

1. Introduction

Evidence-Based Policy (EBP) is premised on the idea that policy effects estimated in one population can be used for predicting the effects of similar, future interventions in distinct target populations. The extrapolation of causal effects to eventual targets, however, remains riddled with substantive challenges. This is because populations often differ in causally relevant respects, so assuming that the effect of an intervention will be the same in the target as in an experiment is typically not justified – this would be extrapolation based on hope (cf. Cartwright, 2012; Steel, 2008).

In the non-structural econometrics literature, *interactive covariate-based approaches* have been proposed to help overcome problems of extrapolation (Crump, Hotz, Imbens, & Mitnik, 2008; Hotz, Imbens, & Mortimer, 2005; Muller, 2014, 2015). In a nutshell, these approaches consider causally relevant differences between populations in the form of *interactive covariates*, i.e. variables that induce differences in causal effects between individuals and between populations. The aim is to take differences in these variables into account by adjusting the expectation of the effect in the target according to how these variables are distributed there.

A key problem is that learning which variables to adjust for, and how these variables are involved in producing the effects of interest will often be exceedingly demanding, e.g. one might need to know whether they play the same causal roles in both populations. This raises concerns about the *extrapolator's circle* (LaFollette & Shanks, 1996; Steel, 2008): the knowledge required about the

CONTACT Donal Khosrowi donal.khosrowi@durham.ac.uk

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

target population may be so extensive that one could learn the effect of interest based on information about the target *alone*. This would make the act of extrapolating *from* an experimental population redundant.

Not all is lost, as there are other strategies that could help underwrite extrapolation. Of particular interest is Steel's (2008) *comparative process tracing* strategy (CPT) that supposedly evades the extrapolator's circle. Although promising, I argue that CPT has trouble evading the extrapolator's circle in many cases of interest in EBP and econometrics. Specifically, I offer a distinction between two kinds of extrapolation, *attributive* and *predictive*, the latter being prevalent in EBP and econometrics. *Attributive extrapolation* aims to attribute *observed effects* causally to their suspected causes. Here, both the intervention of interest as well as its suspected effects have been observed in the target. In contrast, *predictive extrapolation* aims to predict the *future effects* of yet unobserved interventions, so neither the intervention nor its suspected effects have been observed in the target at the time of extrapolation. I argue that in predictive extrapolation the kind of evidence favored by econometricians, i.e. quantitative observational[1] data, cannot be used to tell whether populations are sufficiently similar to license extrapolation. I conclude that this provides a strong case for thinking that EBP researchers and econometricians should consider alternative kinds of support for underwriting extrapolation, including qualitative evidence.

Let me briefly outline the general challenges involved in extrapolation and the distinctive challenge posed by the extrapolator's circle.

2. Extrapolation: the basics

Following Steel (2008), two challenges must be met by any persuasive strategy for extrapolation. First, since causally relevant differences between experimental and target populations will almost always obtain, a persuasive strategy for extrapolation should tell us how extrapolation can proceed successfully despite such differences. It is useful to distinguish between three levels at which such differences can obtain.

The first concerns differences in the *realizations/distributions of variables*. For instance, a causal effect of X on Y may differ between individuals/populations as a result of differences in the value/distribution of some moderating variable W that induces differences in the causal effect. For instance, when higher values of W induce larger effects for one and the same intervention, and W has a higher mean in population B than in A, then the mean effect of one and the same intervention on X will be larger in B than in A.

The second level concerns differences in the *parameters* associated with the variables that figure in the causal mechanisms governing the effects of interest. For instance, X can be causally relevant for Y in two populations, yet the particular *way in which* X is relevant for Y, e.g. the parameter capturing its marginal effect on Y, can differ between populations. The same applies to the way in which moderating variables affects causal relationships between X and Y. For instance, for one and the same intervention, higher values of W might induce larger effects in A, but smaller effects in B.

Finally, the third level at which causally relevant differences can obtain concerns the *basic structure of the mechanisms*[2] that govern the effects of interest. For instance, X can be causally relevant for Y in A but not in B, e.g. because there is no causal pathway connecting X and Y in B. More generally, differences at the third level concern qualitative features of causal mechanisms; e.g. features of the skeletons of the causal graphs representing the mechanisms in both populations such as whether there is some arrow between a pair of variables or not, what is the direction of the arrow, and what is the functional form association between the variables.[3]

Successful extrapolation hinges on similarities and differences between populations at different levels; and if differences at any of these levels obtain, we will need to learn about them and take them into account. This is not an easy task, however, as learning what the relevant similarities and differences are can be extremely challenging.

Moreover, even if we could be successful in identifying them, there is a second challenge to be addressed: the *extrapolator's circle* (LaFollette & Shanks, 1996; Steel, 2008). Specifically, the knowledge

about the target required for extrapolation must not be so extensive that we can identify the causal effect in the target based on information from the target *alone*.[4] This would render the act of extrapolating *from* the experiment redundant. For instance, suppose X causes Y in A and our aim is to determine whether this also holds true in B. Suppose the mechanism in A is $X \rightarrow Z \rightarrow Y$. In order to decide whether X causes Y in B it seems important to learn whether a similar mechanism is instantiated there. However, learning that $X \rightarrow Z \rightarrow Y$ in B makes the extrapolation from A redundant, as one can already answer whether X causes Y in B from information about B alone. This poses a challenge for any persuasive account of extrapolation: any such account should help us extrapolate given only *partial* information about the target, i.e. information that does not, by itself, permit identification of the effect of interest (Steel, 2008, p. 87).

With this background in place, let me expand on strategies for extrapolation offered in the econometrics literature and explain why they fall prey to the extrapolator's circle.

3. Extrapolation in econometrics

Interactive covariate-based strategies for extrapolation offered in the non-structural[5] econometrics literature (Hotz et al., 2005; Muller, 2014, 2015) acknowledge that there are likely to be important differences between experimental and target populations. To overcome these obstacles, they propose a way in which we can still successfully extrapolate by taking such differences into account. In doing so, they focus on differences at the first of the three levels outlined above: differences in the distributions of *interactive covariates*, i.e. variables W that can induce differences in causal effects between individuals and between populations.

Taking such differences into account is a two-step procedure: the first step is to identify a causal effect in an experimental population as a *conditional average treatment effect* (*CATE*, see Muller, 2014, 2015), i.e. an average treatment effect that is estimated conditionally upon the experimental populations' distribution of interactive covariates. This is supposed to capture how the average treatment effect in the experimental population hinges on that population's specific distribution of interactive covariates. The second step is to reweight *CATE* according to the observed distribution of interactive covariates in the target.

For instance, suppose we learn that the causal effect of reducing class sizes X on increasing student performance Y depends on the level of teacher quality W; the higher W, the larger the effect of X on Y. Suppose we want to extrapolate this effect from A to B. Let us assume that the distribution of W in A exhibits a high mean; teachers in A are of high quality. Facing potential differences in W between populations, successful extrapolation requires that we take the distribution of W in B into account. If, for instance, W is lower in B, then the effect to be expected in B will be smaller. To help us predict how much smaller, the *CATE* obtained in A is reweighted according to the observed distribution of teacher quality in B. The approach offered by Hotz et al. (2005) hence aims to permit quantitative extrapolation of causal effects in the presence of causally relevant differences by using suitably identified *CATEs* from an experiment and quantitative observational data on the distributions of relevant interactive covariates in the target.

I will not expand here on the assumptions required for this strategy to proceed successfully; these are laid out together with a proof of the relevant extrapolation theorem in Hotz et al. (2005) and several general concerns about these assumptions are discussed by Muller (2014, 2015). My concern here is different, and more fundamental: interactive covariate-based approaches are only concerned with the first level at which causally relevant differences between populations can obtain, i.e. differences in interactive covariate distributions. However, accounting for these differences is only useful if there are no relevant differences between populations at other levels (this is suggested, but not elaborated in Deaton & Cartwright, 2016, p. 39).

For instance, supplementary teaching S might be positively relevant for the effect of schooling X on student performance Y in an experimental population; it helps students review material discussed in class. However, it might be negatively relevant in a target where students are stigmatized by their

peers for being in need of supplementary teaching, making them less confident in their abilities and decreasing their performance on tests.[6] Clearly, adjusting for differences in *S* is only useful if we are confident that populations do not differ at any of the lower levels; specifically not with respect to *whether* and *how S* is involved in producing the effect of interest. Moreover, merely *assuming* that populations are similar at lower levels would amount to extrapolation based on hope. Populations frequently differ in their structural makeup, e.g. because institutions, norms, individuals' psychological characteristics, and other features differ between them. So we need to support empirically the claim that populations are sufficiently similar to warrant extrapolation. That is not only difficult, however, but even if feasible, raises concerns about the extrapolator's circle.

Let me start at the most basic level to illustrate. To support the assumption that experimental and target populations are similar with respect to the basic structure of causal mechanisms we need to learn something about the mechanisms in both populations. Say, for instance, the mechanism in the experimental population is understood to be $X \rightarrow Z \rightarrow Y$. Then, in order to ensure that the mechanism in the target is similar, we might need to learn whether all causal relations comprising this mechanism are present there as well, i.e. we need to learn that $X \rightarrow Z$ and that $Z \rightarrow Y$. But learning this makes the information obtained from the experimental population redundant to answering whether *X* is causally relevant for *Y* in the target.

Similar concerns apply when learning about similarities in *parameters* associated with variables that figure in the mechanisms. While we might be able to learn how a variable *W* induces differences in a causal effect in an experimental population, this only gives us half the information we need. To see this, let us assume that we are fortunate enough know the structural equations governing individuals' outcomes. Let *Y* be the outcome, *X* the treatment variable, β the parameter that captures the 'pure' causal effect of *X*, *W* an interactive covariate of the $X-Y$-effect, γ the parameter that captures the marginal effect of *W* on the causal effect of *X* on *Y*, and *v* an idiosyncratic error capturing the effects of other variables on *Y*. For ease of illustration, let the outcome equations for individuals in population *A* and *B* be of an additively separable form:[7]

$$Y_A \Leftarrow c + \beta * X_A + \gamma_A * W_A * X_A + v_A$$

$$Y_B \Leftarrow c + \beta * X_B + \gamma_B * W_B * X_B + v_B$$

It is easy to see that the marginal effect on *Y* induced by a given change in *X* depends on β, γ and *W*, since $\Delta Y / \Delta X = \beta + \gamma * W$. Now, even if β is the same between populations, adjusting for differences in *W* can only proceed successfully if $\gamma_B = \gamma_A$, i.e. if the way in which *W* induces differences in the effect is the same in both populations. So even if we are fortunate enough to have learned γ_A, we still need to learn γ_B in order to validate that $\gamma_B = \gamma_A$ (or is otherwise sufficiently similar).

Just like validating that mechanisms are similar between populations, establishing that they are similar in population-level parameters is generally difficult. For once, observational data for estimating γ_B might not be available. Even if they are, *unbiased* estimation of γ_B requires substantive assumptions, e.g. that there are no common causes of *W* and differences in treatment effects that could induce significant, but ultimately spurious interactions between *X* and *W*. To avoid such assumptions, γ_B can be identified by performing *factorial experiments* where both *X* and *W* are exogenously varied (see e.g. Imai, Tingley, & Yamamoto, 2013; Pearl, 2014). However, doing so may not only be difficult – think of variables such as age that cannot be meaningfully intervened on – but also, factorial experiments in the target involve intervention on *X* and hence trivially fall prey to the extrapolator's circle as we can learn the causal effect of interest by doing so.

Proponents of interactive covariate-based extrapolation may object at this point that causally relevant differences at the levels of parameters and basic structure of mechanisms pose no special problem to their approach and can be handled if there are observable proxy variables that correlate with these differences, so we can simply adjust for differences in the distributions of such proxy variables.

There are three reasons to be sceptical about this possibility. First, it is unclear whether differences in parameters and the basic structure of causal mechanisms frequently have readily observable correlates that are amenable to this strategy. Put simply, agents and populations do not typically wear mechanism-types or parameter values on their sleeves. Second, features of causal mechanisms (and differences in such features) that are important for extrapolation often do not readily manifest themselves in ways other than in agents' behavioral response to the specific intervention of interest; think for instance about latent psychological characteristics. So while there could be proxy variables that correlate with important parameters and features of mechanisms, measurements of such variables would often only be useful after the intervention of interest was already experienced by agents, thus again raising concerns about the extrapolator's circle. Third, even if many causally relevant differences in parameters and the basic structure of mechanisms had readily observable proxies to permit interactive-covariate-based reweighting, measuring and accommodating differences in such variables would still require extensive causal knowledge about both populations, including details concerning how proxy variables are associated with the underlying differences in parameters and basic causal structure and whether this association is the same in both populations. This suggests that the concerns developed here are not easily remedied by appealing to the principled possibility of adjusting for differences in proxy variables; at least not without substantive assumptions and raising additional concerns about the extrapolator's circle.

The upshot is this: interactive covariate-based extrapolation proceeds on the assumptions that experimental and target populations are relevantly similar at the level of the basic structure of causal mechanisms as well as the parameters associated with the interactive covariates by which adjustment proceeds. Validating these assumptions, however, will not only often be difficult, but is also likely to fall prey to the extrapolator's circle.

This suggests the need for a supplementary strategy to underwrite interactive covariate-based extrapolation; one that evades the extrapolator's circle. At least one such strategy has been proposed by Steel (2008). Let me briefly outline Steel's proposal before I proceed to argue that, although promising, it will not help EBP researchers and econometricians evade the extrapolator's circle in the kinds of cases they routinely encounter, and given their preference for using quantitative observational data from the target to facilitate extrapolation.

4. Steel's comparative process tracing

According to Steel (2008), a strategy called *comparative process tracing* (CPT) can help us extrapolate causal relevance claims, e.g. whether *X* is causally relevant for *Y* in *B* if it is causally relevant in *A*, while evading the extrapolator's circle. In virtue of this, CPT could be a promising candidate to underwrite interactive covariate-based extrapolation by helping us determine whether populations are similar at the level of the basic structure of causal mechanisms.

Steel's strategy proceeds in two steps. The first is to learn the mechanism in the experimental system by means of *process tracing*, i.e. specifying distinctive marks one would expect to be present if, and preferably only if, a putative mechanism is operating, and determining empirically whether these distinctive marks are present or absent, thus speaking for or against the putative mechanism being operational (see e.g. Beach, 2017; Salmon, 1984, ch. 4).

The second step is to *compare* mechanisms in the experimental and target populations at so-called *downstream bottleneck* stages, i.e. stages of the mechanism that are closest to the outcome and through which any effect of an intervention must be transmitted (Steel, 2008, p. 89).

To illustrate, Steel offers the example of extrapolating the carcinogenicity of Aflatoxin B1 from animals to humans. Here, the carcinogenicity of Aflatoxins (a class of spoilage mold metabolites), including a particularly potent type called Aflatoxin B1 (AFB1), was experimentally established in a variety of animal species. These carcinogenic effects were also learnt to vary significantly between species as a function of between-species differences in metabolic mechanisms. Together with quantitative observational evidence from human populations indicating

significantly higher incidence of hepatocellular carcinoma (HCC) for humans that had been exposed to AFB1, this raised the extrapolative query whether humans are similarly susceptible to these effects. Straightforward causal attribution of observed HCC to AFB1 exposure in humans was precluded, however, since observed populations also exhibited high rates of Hepatitis B virus infection (HBV), a predominant cause of HCC in humans. Hence, an important step in clarifying the carcinogenicity of AFB1 in humans was to investigate whether a similar metabolic mechanism that governs the carcinogenicity of AFB1 in animals is also operating in humans. In order to evade the extrapolator's circle, these comparisons of mechanisms in animals and humans needed to proceed without learning the full mechanism in humans. According to Steel, CPT helped achieve this in the following way (but see Reiss, 2010 who questions whether the extrapolator's circle was indeed evaded):

Let *C* be exposure to AFB1, *Y* the cancer outcome, and *X, A, Z, B, E* the intermediate stages of the mechanism involved in metabolizing AFB1 in animals:

$$C \rightarrow X \rightarrow A \rightarrow Z \rightarrow B \rightarrow E \rightarrow Y$$

Suppose the stages at which mechanisms between animals and humans are most likely to differ are *Z* and *E*, where *E* is a bottleneck through which any influence of AFB1 on the outcome is transmitted. Then, to ensure that populations are similar with respect to mechanisms, it is sufficient to compare mechanisms at *E*, a descendant of *C* and the nearest parent of *Y*. If mechanisms are similar at this stage, in the sense that variation induced in *C* is transmitted up to *E*, then extrapolation of causal relevance claims can proceed successfully even without full knowledge of the mechanism in the target. The reasoning is that if experimental and target systems are relevantly different at any stage upstream of *E* and downstream of *C*, then variation induced in *C* will not transmit up to *E*. Conversely, if variation induced in *C* does transmit up to *E*, then either the mechanisms are similar between *C* and *E*, or they are dissimilar but the dissimilarities are not relevant since they do not curtail the transmission of variation from *C* to *E*.

Against this background, Steel's extrapolation theorem says that using CPT to learn that mechanisms in both populations are similar at downstream bottleneck stages helps us to successfully extrapolate claims of causal relevance without falling prey to the extrapolator's circle. (Steel, 2008, p. 113).

At face value, this result could be useful for underwriting interactive covariate-based extrapolation as it seems to offer a way to evade the extrapolator's circle at least when it comes to supporting that experimental and target populations are similar at the level of mechanisms.

In what follows, I offer reasons to think that CPT will not be useful for this purpose in the kinds of cases that econometricians and EBP researchers typically encounter, and given their preference for using quantitative observational evidence from the target to facilitate extrapolation. To help delineate these cases, I develop a distinction between two kinds of extrapolation: *attributive* and *predictive*.

5. The limits of comparative process tracing

The extrapolation Steel discusses in his AFB1 example is of a specific kind. This *attributive extrapolation* aims to attribute an observed effect causally to its suspected causes. The kind of extrapolation typically encountered in EBP and econometrics is importantly different. This *predictive extrapolation* aims to predict the future effects of (interventions on) suspected causes. This proceeds under conditions where neither the intervention of interest nor its suspected effects have yet been observed in the target. This distinction is related to the familiar distinction between investigating the effects of causes and the causes of effects. In the present case, we are interested however not in investigating what the effects of causes or causes of effects are in a particular study setting, but rather with reaching extrapolative conclusions about the causes of observed effects, or about the effects of future interventions on suspected causes in a distinct target.

In what follows I argue that problems of predictive extrapolation are unlikely to be overcome by using a combination of CPT and quantitative observational evidence without falling prey to the extrapolator's circle.

To illustrate, let me invoke Steel's AFB1 extrapolation case again. One distinctive feature of this extrapolation is that it proceeded against the background of observational evidence on humans where both the suspected cause of interest, AFB1 exposure, as well as its suspected effect, HCC, were jointly realized and observed. The primary aim of the extrapolation was not to answer whether AFB1 is a cause of HCC in humans *simpliciter*, but rather whether AFB1, *rather than* HBV, or *in addition to* HBV, was a cause of *observed* HCC in humans that *had been exposed* to AFB1. In more general terms, the extrapolative query was whether *X*, rather than, or in addition to *Z*, is a cause of *observed effects* on *Y* in the target. I call this type of extrapolation *attributive extrapolation*.

This is importantly different from *predictive* extrapolation, which is typically encountered in EBP and econometrics contexts. Here, the point of departure is usually a causal effect of an intervention identified in an RCT or quasi-experimental study and the question is what the effects of the same (or a similar) intervention will be in a novel target. What is distinctive of these cases is that neither the intervention on *X* nor its suspected effects on the outcome *Y* have yet been observed in the target. The reason is simply that most studies in EBP and econometrics investigate the effects of *novel* interventions, so it may usually be assumed that these interventions have never been historically experienced in envisioned policy targets at the time of extrapolation. This means that the aim of extrapolation is typically *predictive* rather than attributive: will *yet unobserved interventions on X* cause *future, yet unobserved* effects on *Y* in the target?

This difference between attributive and predictive extrapolation goes hand in hand with important differences in the evidence that is available to support extrapolation. Steel's aim is to highlight the importance of process tracing evidence, i.e. observations of the presence or absence of distinctive marks of a hypothesized mechanism.

It is important to recognize, however, that CPT in Steel's case proceeded against the background of several other kinds of supplementary evidence that established a basis for comparing mechanisms, including (1) evidence that helped to characterize the mechanism of interest in animals, (2) evidence indicating between-species variation in the effect of interest that could present obstacles to extrapolation from animals to humans, as well as (3) evidence concerning between-species differences in mechanisms that induced these between-species differences in effects.

As Steel recognizes, an immediate problem in many social science contexts is that such evidence is difficult to produce (2008, Ch. 8). Concerning (1), mechanisms governing the outcomes of interest are often difficult to observe even in the experimental population; think for instance about the psychological mechanisms involved in governing individuals' economic choice behaviors. This is aggravated by the extant emphasis in EBP and econometrics to focus on studying the effects of causes rather than (also) investigating the causal mechanisms governing these effects, meaning that evidence with bearing on questions of mechanistic similarity and difference is typically not considered relevant (see e.g. Kern, Stuart, Hill, & Green, 2016), and only few attempts are made to produce such evidence (see Cook, 2002; Deaton, 2009; Heckman & Smith, 1995 for related criticisms; but see Imai et al., 2013 for an exception). Concerning (2), observational evidence concerning differences in causal effects is somewhat easier to produce (see e.g. Athey & Imbens, 2016 for machine learning methods to detect heterogeneous causal effects). Yet, attributing such differences to underlying mechanistic differences is again difficult for lack of (3), i.e. evidence indicating which differences in mechanisms induce such differences in effects.

However, even if such evidence could be readily obtained from experimental populations, to *compare* mechanisms we still need some evidence characterizing the mechanism in the target. This is crucial. In contrast to the AFB1 example where experiments on components of the putative mechanism in the target could be performed (e.g. on human cell cultures), in many contexts of interest in EBP and econometrics similar means for observing mechanisms in the target are unavailable (Steel, 2008, p. 166). Instead, the most salient way in which EBP researchers and econometricians

could compare mechanisms in line with standard methodological tenets and evidential preferences in these fields is by using quantitative observational data from the target. Such data could help determine whether distinctive marks of the suspected mechanism, e.g. in the form of distinctive covariance and (conditional) probabilistic dependence/independence signatures between variables, are realized there.[8]

In Steel's example, such evidence is supplied by observational studies on humans. These studies offered covariance information suggesting that relative risk of HCC is significantly higher in humans that have been exposed to AFB1 and that this association remains stable even when conditioning on HBV infection (IARC, 1993).

Such evidence is not available in predictive extrapolation cases. The crucial difference here is that when neither the intervention of interest nor its suspected effects have yet been observed in the target, quantitative observational evidence cannot speak to questions about mechanistic similarity and difference.

An example helps illustrate this. Suppose we learn that distributing free insecticide-treated bed nets helps decrease Malaria infection rates in population *A*. Suppose further that bed nets must be properly installed to curtail Malaria infection, and that whether nets are in fact properly installed can differ significantly between populations, e.g. agents in some populations might use them as fishing nets instead (McLean et al., 2014). Let me represent this by the simplistic mechanism $X \rightarrow Z \rightarrow Y$, where Z, the number of properly installed nets, is a mediating variable on the path from distributed nets X to Malaria infection Y (where Z is negatively relevant for Y).

How can we make sure that this mechanism is sufficiently similar between an experimental population *A* and a novel target *B* where bed nets have not yet been distributed? Quantitative observational evidence that could help indicate that the mechanism in *B* is similar to that in *A* would be that Z is higher conditional on X than unconditionally, indicating that distributed nets are properly installed, and that Y is lower conditional on Z than unconditionally, suggesting that properly installed nets in fact reduce Malaria infection.[9]

The crucial problem is that such information cannot be obtained from the target if no bed nets have ever been distributed there. If that is the case then X and Z exhibit no variation, since $X = 0$, and $Z = 0$ for all individuals, and Y will only assume its natural value that is induced through relevant Malaria infection pathways.[10] This means that Z conditional on X and Z unconditionally will be equal, and that Y conditional on Z, and Y unconditionally will be equal as well. So if no bed nets have ever been distributed in the target, there will be no (co-)variation in the outcomes of interest or the intermediate stages of the mechanism that could help us tell whether mechanisms are sufficiently similar.

More generally, in cases where the intervention of interest has never historically been experienced and observed in the target, quantitative observational data on variables that figure in the suspected mechanism are not informative about mechanistic similarities and differences between populations. This means that if, as econometricians and EBP researchers do, we primarily consider quantitative observational data from the target relevant for underwriting extrapolation, we cannot tell whether the target exhibits characteristic signatures of the suspected mechanism being operational. As Steel anticipates, '[…] the operation of a program can be examined only where it is implemented […]' (2008, p. 166). So while this problem can be remedied by intervening on X in the target, doing so would trivially fall prey to the extrapolator's circle.

To be sure, one could argue that quantitative observational evidence from the target can still have *indirect* bearing on questions of mechanistic similarity and difference if there have been similar, and well-understood interventions (or exogenous changes) in the target in the past.[11] This seems possible, but would also seem to require substantive assumptions concerning how such past interventions (or exogenous changes) relate to those of current interest, e.g. whether their effects are governed by the same mechanisms and in the same way, as well as whether the intervention of current interest is structure-altering or not. Such assumptions are not recognizably weaker than, and similarly difficult to support as those at issue here.

The concerns outlined above are not surprising. Quantitative observational data only have bearing on questions concerning features of causal mechanisms if there is sufficient variation in at least some of the putatively causally relevant variables. More generally, we might say that quantitative observational evidence can only be informative about the causal mechanisms governing observable phenomena if these phenomena have already obtained in some way or another, and have consequently had the opportunity to write, as it were, distinctive marks, symptoms, signatures etc. into the data that we can obtain. Without such opportunity, quantitative observational evidence from the target remains a poor guide to clarifying issues of mechanistic similarity and difference.

6. Predictive extrapolation: where next?

The previous discussion suggests three things. First, interactive covariate-based strategies for extrapolation involve wide-ranging but ultimately unsubstantiated assumptions that populations are sufficiently similar at the level of mechanisms and parameters relevant to the effects of interest. Second, empirically validating these assumptions raises concerns about the extrapolator's circle. Third, even Steel's CPT is not immune to these concerns: at least in predictive extrapolation cases, quantitative observational data from the target are of little help in clarifying issues of mechanistic similarity and difference. So even if econometricians and EBP researchers were to use CPT, they could not rely on quantitative observational data from the target.

This is not to suggest that interactive covariate-based strategies are fundamentally flawed. It is practically difficult, but perhaps not insurmountably so, to support that populations are relevantly similar at the level of mechanisms and parameters. Similarly, my aim is not to suggest that CPT is an inadequate strategy for extrapolation in general. I consider CPT to be a promising strategy; but my concern is that predictive extrapolation poses distinct challenges for CPT, specifically that quantitative observational data from the target are not useful for CPT in such cases. So if preferences for such evidence are maintained, it seems that interactive covariate-based extrapolation cannot be underwritten by CPT, and if applied at all, would need to proceed on hope that populations are sufficiently similar rather than evidence that this is so.

I now want to consider some suggestions for what might be done about this. Since the primary aims of this paper are critical in nature, I will not attempt to develop an alternative strategy for underwriting extrapolation here. However, it seems useful briefly to consider some ways to respond to the challenges put forward; specifically proposals offered in EBP-related literatures that are not married to preferences for quantitative data.

There is a rich literature on supporting process tracing, including for purposes of extrapolation, by *qualitative* social scientific evidence, e.g. sociological, anthropological, and ethnographic evidence obtained from sources such as interview studies, participatory observation, and expert judgment (see e.g. Blatter & Blume, 2008; Kay & Baker, 2015; Schmitt & Beach, 2015; see Fairfield & Charman, 2017 for a Bayesian approach to integrating different kinds of qualitative evidence). Econometricians and EBP researchers have so far been reluctant to consider such evidence (see Kern et al., 2016). However, the arguments provided here suggest that this reluctance is misguided, and that producing, using, and integrating other kinds of evidence, including qualitative evidence, may be useful, and in some cases perhaps even necessary, for underwriting extrapolation beyond the level of mere hope that crucial assumptions are satisfied.

Let me draw on the bed net example again to illustrate how this could proceed. For instance, analogously to in-vitro studies on human cell cultures that helped pin down specific features of parts of the suspected mechanism in humans in Steel's AFB1 example, it might be possible to investigate whether the insecticidal effects of the bed nets to be distributed in the target do in fact obtain on Mosquitos sampled from the target; at least qualitatively. So for some parts of the suspected mechanisms in the target, it may be relatively straightforward to isolate and study parts of it (or well-understood analogues of it), without (1) having to rely on quantitative observational data from the target

(such as observing that infection rates are lower conditionally on properly installed nets than unconditionally) or (2) introducing bed nets in the target and triggering the extrapolator's circle.[12]

For other causal relations this may be more challenging. For instance, it seems that investigating whether agents in the target will properly install bed nets can again raise concerns about the extrapolator's circle, e.g. when we introduce bed nets provisionally in (at least part of) the target. Proponents of qualitative approaches to EBP might suggest that this could be avoided by considering sources of evidence such as participatory observation, agents' self-reports, or expert judgment. For instance, they might point out that agents can sometimes reliably report on counterfactual states of affairs that have bearing on questions of mechanistic similarity and difference; e.g. when these counterfactual states are importantly determined by agents' own decision-making (see e.g. Fairfield & Charman, 2017; Kay & Baker, 2015; for such suggestions). If, for instance, agents do in fact have alternative uses for bed nets, such as using them as fishing nets, then it seems that they could, under some conditions, reliably report that, if they were given free bed nets, they would not use them as bed nets but rather as fishing nets. Similarly, agents may also be able to report on counterfactual states of affairs that hinge on other agents' decision making, or on existing social norms, where agents, as well as experts with local knowledge of the target, might be able to anticipate, at least qualitatively, how these norms may interact with the intervention of interest (see Cartwright & Hardie, 2012 for related suggestions).[13]

Of course, considering qualitative evidence also raises a host of new challenges. Agents may be overconfident about their propensity to adhere to implementation protocols; they may experience substantial difficulties in anticipating the effects of some interventions, such as predicting the effect of deworming on student achievement; they may be incentivized to strategically misreport their expected future behaviors, and so on.

More generally, agents' self-reports are often plausibly suspected to be unreliable, and various important precautions need to be undertaken to support the reliability of such evidence, e.g. triangulating qualitative conclusions by considering multiple sources and using different elicitation methods, having agents report on others' behaviors instead of their own, ensuring that agents are not improperly incentivized to strategically misreport, etc. (see Schmitt & Beach, 2015 for practical examples concerning the importance of such precautions).

What is more, it is not to be expected that qualitative evidence of the kind outlined above will be sufficiently informative by itself to tell us how much exactly, for instance, a particular effect will be amplified or diminished by local causal features of the target. However, it is important to recognize that this does not preclude quantitative predictions of causal effects in the target. If qualitative evidence increases our confidence that crucial features of causal mechanisms are qualitatively similar between populations, e.g. that a moderating variable W is likely to interact with an intervention in the same qualitative way in both populations, this can offer important support (although perhaps not full-fledged warrant) for the assumptions that are necessary for interactive covariate-based extrapolation to proceed. If this is successful, interactive covariate-based strategies may justifiably be used to obtain *quantitative* predictions of causal effects, just as envisioned by their proponents.

So qualitative evidence is not a silver bullet to address the shortcomings of interactive covariate-based strategies. However, when quantitative data from the target are unlikely to help clarify whether populations are sufficiently similar to license extrapolation at all, considering qualitative evidence with bearing on these issues, despite its potential shortcomings and the additional methodological burden placed on us, may still be recognizably superior to proceeding on mere hope that populations are sufficiently similar.

Most importantly, considering qualitative evidence promises to help us evade the extrapolator's circle. Here, the intervention is not introduced in the target, but only hypothetically in the minds of agents who may possess relevant expertise to report on features of causal mechanisms and processes that they are part of, and that have important bearing on the effects of interest. This would steer clear of the extrapolator's circle because while agents' self-reports may help us rule out

important causally relevant differences between populations, the casual effects of interest could probably not be learned by asking them any number of questions. So qualitative evidence can be useful for clarifying issues of similarity and difference between populations, but is not a sufficient means to predict causal effects in the target.

So what is the main suggestion for how extrapolation in econometrics and EBP should proceed in light of the arguments provided here? It seems clear that predictive extrapolation (which is predominant in econometrics and EBP) poses distinct challenges for interactive covariate-based strategies (as well as for CPT). While there is no obvious remedy, it seems that considering qualitative evidence with bearing on questions of similarity and difference between populations is an option that should be explored in more detail, as such evidence might be able to give us at least some purchase on whether the assumptions required for interactive-covariate-based extrapolation are satisfied.

In light of this, it seems reasonable to suggest that the ability of qualitative evidence to speak to these issues should be investigated further, and that the production and use of such evidence should be encouraged in widely circulated methodological guidelines such as those issued by the Campbell Collaboration, the What Works Clearinghouse, CONSORT, GRADE, J-PAL and others. In addition, more attempts should be made to develop strategies for *integrating* quantitative and qualitative evidence in domain-general theories (sometimes called program theories, theories of change, or logic frames) that aim to offer comprehensive accounts of *how* the interventions of interest achieve their intended effects, and under what conditions they might fail to do so (see e.g. Davey et al., 2017 for similar suggestions). Some EBP institutions such as J-PAL, 3ie and others already make attempts along these lines (White, 2009). These are only early steps however, and a persuasive, general methodology for underpinning extrapolation by means of integrating qualitative and quantitative evidence is still missing.

As the arguments provided here suggest, there is much promise in developing such a methodology. It seems that there can be cases where qualitative and quantitative evidence, when considered in tandem, can help us extrapolate causal effects in a way that is superior to doing so based on either type of evidence alone. Qualitative evidence by itself can at best clarify issues of qualitative causal relevance. Quantitative evidence by itself, on the other hand, can help us make extrapolative predictions of causal effect magnitudes in the target, but these predictions are only credible if crucial assumptions about similarities between populations are empirically supported. At least in predictive extrapolation it is clear that this role cannot be played by quantitative observational evidence from the target. But as the arguments provided here suggest, qualitative and quantitative evidence can play complementary roles: one helps clarify whether populations are similar at the level of basic causal structure, the other helps investigate causal effect magnitudes of interventions implemented in one setting, and with adjusting for differences in the distributions of variables that can modify these effects. Considered together, both types of evidence can hence underwrite extrapolative conclusions that would not be accessible from either type of evidence alone.

Providing the details of a methodology for integrating qualitative and quantitative evidence is beyond the scope of this paper, but I hope that the arguments provided here will reinforce similar suggestions made by other philosophers (e.g. Cartwright, 2013; Cartwright & Hardie, 2012; Grüne-Yanoff, 2016) by providing reasons to think that extrapolation may not only be greatly facilitated by considering qualitative evidence pertaining to the causal mechanisms governing the effects to be extrapolated, but that a wide range of real-world extrapolations may be exceedingly difficult to underwrite without doing so. This, I hope, will help motivate further contributions that encourage econometricians and EBP researchers to add previously neglected kinds of evidence to their arsenals in the pursuit of underwriting extrapolation by more than hope alone.

Notes

1. For the present purposes I understand the distinction between observational and experimental in the standard sense that observational approaches are passive and that experimental approaches either involve intervention by the investigator, or exploit variation in the treatment variable that is justifiably believed to be exogenous.

2. By *mechanism* I mean a set of causal relationships (or representation thereof) between variables that underpin and govern observable phenomena, such as transmitting the effects of interventions on one variable to another. I do not wish to make more precise commitments, as they would detract from the generality of my arguments.
3. One might argue that we can represent the absence of causal relations between X and Y already at the second level, e.g. by setting the parameter for such relations zero. While this is possible, I prefer to keep two issues distinct: quantitative issues concerning the signs and magnitudes of marginal causal effects should be represented at the level of parameters, whereas qualitative issues concerning the presence, absence, and direction of causal relations should be represented at the level of the basic structure of the mechanisms. More generally, the three levels distinguished here are mostly of pragmatic value and do not constitute a substantive commitment that they are genuinely distinct levels of causal analysis, nor that it is ever straightforward to classify features that induce causally relevant differences as features that obtain at either one of the levels but not others.
4. A more nuanced, gradual version of this concern would be that the information from the experiment is *almost* redundant to the extrapolation, e.g. when even major variations in experimental results would have negligible impact on our quantitative predictions of the effect in the target.
5. For the present purposes I focus on non-structural microeconometrics, specifically the so-called treatment-effects literature, which is concerned with estimating and extrapolating causal effects from experimental and quasi-experimental data. For detailed examinations of the differences between, and history of, structural and non-structural, reduced-form approaches see Boumans (2005); for a critical discussion see Keane (2010).
6. One might argue that rather than the parameter associated with W being positive in A but negative in B, what happens here is rather that W is associated with the outcome in B via an additional causal pathway that is mediated by stigmatization, which is negatively relevant for performance. I am open to alternative characterizations, as not much hinges on this. The point is merely that we need to ensure that there are no differences in whether (causal structure) and how (parameters) W is involved in producing Y; otherwise, adjusting for differences in the value/distribution of W will not help us predict the correct casual quantity in B.
7. The equations simply encode the causal assumptions elaborated here and represent that Y is causally determined in accordance with the equations. Indices are suppressed for simplicity, which means that we assume individuals within A and B respectively to be perfectly alike. Importantly, the particular functional form here, where the term involving W is additively separable, is only assumed for ease of illustration. The problem highlighted persists in the more general case where $Y \Leftarrow f(\gamma, W, X, v)$. So long as $f(\cdot)$ involves some interaction between X and W, where the marginal effect of W is not separable from X, we need to ensure that the functional form of $f(\cdot)$ and the value of γ are the same between populations.
8. This seems coherent with how extrapolation is supposed to proceed according to Hotz et al. (2005). Recall that, on their approach, the evidence from the target used to facilitate extrapolation is quantitative observational evidence concerning the distributions of putatively relevant variables that may induce differences in causal effects. Exemplary applications such as Dehejia, Pop-Eleches, and Samii (2015) and Gechter (2016) suggest that econometricians and EBP researchers would also be inclined to rely on quantitative observational data from the target to support that experimental and target populations are relevantly similar at the levels of parameters and mechanisms.
9. This is called *pattern evidence* on Beach and Pedersen's (2016) typology of process tracing evidence.
10. In the macroeconometrics literature, this is known as the problem of *non-excitation* (cf. Salmon & Wallis, 1982; Engle, Hendry, & Richard, 1983).
11. Steel anticipates this general intuition when discussing concerns surrounding structure-altering interventions (2008, pp. 157–160). I thank Wendy Parker for calling my attention to this.
12. I thank an anonymous referee of this journal for suggesting this example.
13. This is not dramatically different from what Steel's CPT recommends. The arguments presented here add an important nuance, however, which is that in predictive extrapolation, for lack of observational evidence indicating that there is *some* causal pathway from the intervention variable to the outcome of interest, CPT may need to be supported with *more* evidence that has bearing on questions of mechanistic similarity. When available evidence from the target is insufficient to clarify these questions, this may, again, make it likely that we fall prey to the extrapolator's circle. At least in these cases, the attributive/predictive extrapolation distinction has important ramifications for CPT: it can evade the extrapolator's circle in many, but perhaps not in all cases.

Acknowledgments

I would like to thank two anonymous referees of this journal, as well as Julian Reiss, Nancy Cartwright, Wendy Parker, and members of the CHESS and K4U research groups at Durham University for their many valuable comments and

suggestions on earlier versions of this paper. Moreover, I would like to thank the audiences at INEM2017 and EPSA2017 for raising several important points that helped improve the arguments developed here. My work on this paper was financially supported by an AHRC Northern Bridge Doctoral Studentship (grant number: AH/L503927/1), a Durham Doctoral Studentship, and a Royal Institute of Philosophy Jacobsen Studentship. I am very grateful for this support.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Donal Khosrowi is currently a doctoral candidate in Philosophy at the Centre for Humanities Engaging Science and Society at Durham University. His doctoral research focuses on strategies for extrapolating causal effects. His broader research interests and recent publications are concerned with causal inference in social science, scientific representation, and values in science.

ORCID

Donal Khosrowi http://orcid.org/0000-0002-9927-2000

References

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Beach, D. (2017). Process-tracing methods in social science. *Oxford Research Encyclopedia of Politics*. Retrieved from http://politics.oxfordre.com/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-176

Beach, D., & Pedersen, R. B. (2016). *Causal case studies: Foundations and guidelines for comparing, matching and tracing*. Ann Arbor: University of Michigan Press.

Blatter, J., & Blume, T. (2008). In Search of co-variance, causal mechanisms or congruence? Towards a plural understanding of case studies. *Swiss Political Science Review*, *14*(2), 315–356.

Boumans, M. J. (2005). *How economists model the world into numbers*. London: Routledge.

Cartwright, N. D. (2012). Presidential address: Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science*, *79*(5), 973–989.

Cartwright, N. D. (2013). Evidence, argument and prediction. In *EPSA11 perspectives and foundational problems in philosophy of science, the European philosophy of science association proceedings 2*. Basel: Springer.

Cartwright, N. D., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*, 175–199.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics*, *90*(3), 389–405.

Davey, C., Hassan, S., Bonell, C., Cartwright, N., Humphreys, M., Prost, A., & Hargreaves, J. (2017). *Gaps in evaluation methods for addressing challenging contexts in development* (CEDIL PreInception paper). London.

Deaton, A. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. *Proceedings of the British Academy, 2008 Lectures*, *162*, 123–160.

Deaton, A., & Cartwright, N. (2016). *Understanding and misunderstanding randomized controlled trials* (Technical report). National Bureau of Economic Research.

Dehejia, R., Pop-Eleches, C., & Samii, C. (2015). Local to global: External validity in a fertility natural experiment. *SSRN Electronic Journal*. doi:10.2139/ssrn.2647649

Engle, R. F., Hendry, D. F., & Richard, J.-F. (1983). Exogeneity. *Econometrica*, *51*(2), 277–304.

Fairfield, T., & Charman, A. (2017). Explicit Bayesian analysis for process tracing: Guidelines, opportunities, and caveats. *Political Analysis*. ISSN 1047-1987.

Gechter, M. (2016). *Generalizing the results from social experiments: Theory and evidence from Mexico and India* (Unpublished manuscript). Pennsylvania State University.

Grüne-Yanoff, T. (2016). Why behavioral policy needs mechanistic evidence. *Economics and Philosophy*, *32*(3), 463–483.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*, 85–110.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, *125*, 241–270.

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*, 5–51.

International Agency for Research on Cancer. (IARC). (1993). Some naturally occurring substances: Food items and constituents, heterocyclic aromatic amines and mycotoxins. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 56. Lyon.

Kay, A., & Baker, P. (2015). What can causal process tracing offer to policy studies? A review of the literature. *Policy Studies Journal*, *43*(1), 1–21.

Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, *156*, 3–20.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127.

LaFollette, H., & Shanks, N. (1996). *Brute science: Dilemmas of animal experimentation*. New York: Routledge.

McLean, K. A., Byanaku, A., Kubikonse, A., Tshowe, V., Katensi, S., & Lehman, A. G. (2014). Fishing with bed nets on Lake Tanganyika: A randomized survey. *Malaria Journal*, *13*, 395. doi:10.1186/1475-2875-13-395

Muller, S. M. (2014). *Randomised trials for policy: A review of the external validity of treatment effects* (Southern Africa labour and development research unit working paper 127). University of Cape Town.

Muller, S. M. (2015). Causal interaction and external validity: Obstacles to the policy relevance of randomized evaluations. *The World Bank Economic Review*, *29*(1), S217–S225.

Pearl, J. (2014). Reply to commentary by Imai, Keele, Tingley and Yamamoto concerning causal mediation analysis. *Psychological Methods*, *19*(4), 488–492.

Reiss, J. (2010). Review: Across the boundaries: Extrapolation in biology and social science. *Economics and Philosophy*, *26*, 382–390.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Salmon, M., & Wallis, K. F. (1982). Model validation and forecast comparisons: Theoretical and practical considerations. In G. C. Chow & P. Corsi (Eds.), *Evaluating the reliability of macroeconomic models* (pp. 219–249). New York: John Wiley. Chap. 12.

Schmitt, J., & Beach, D. (2015). The contribution of process tracing to theory-based evaluations of complex aid instruments. *Evaluation*, *21*(4), 429–447.

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

White, H. (2009). Theory–based impact evaluation: Principles and practice. *Journal of Development Effectiveness*, *1*(3), 271–284.