2 How to Learn about Causes in the Single Case

Nancy Cartwright

2.1 Introduction

The case study is a broad church. Case studies come in a great variety of forms, for a great variety of purposes, using a great variety of methods – including both methods typically labelled 'qualitative' and ones typically labelled 'quantitative'.[1] My focus here is on case studies that aim to establish causal conclusions about the very case studied. Much of the discussion about the advantages and disadvantages of case study methods for drawing causal conclusions supposes that the aim is to draw causal conclusions that can be expected to hold more widely than in the case at hand. This is not my focus. My focus is the reverse. I am concerned with using knowledge that applies more widely, in consort with local knowledge, to construct a case study that will help predict what will happen in the single case – *this* case, involving *this* policy intervention, *here* and *now*. These involve what philosophers call a 'singular causal claim' – a claim about a causal connection in a specific single individual case, whether the individual is a particular person, a class, a school, a village or an entire country, viewed as a whole. It is often argued that causal conclusions require a comparative methodology. On this view the *counterfactual* is generally supposed to be the essence of singular causality: In situations where treatment T and outcome O both occur, 'T caused O' means[2] 'If T had

[1] For a nice discussion of case study types see Morgan (2014); see Byrne and Ragin (2009) for a good text surveying case-based methods.

[2] Or at least it is supposed that the causal claim is true if and only if the counterfactual is. This has led to endless discussion in philosophy of how to treat putative counterexamples, for example cases of overdetermination and preemption. For further discussion, see Menzies (2014).

Published online by Cambridge University Press

not occurred, then O would not have'.[3] And it is additionally supposed that the only way to establish that kind of counterfactual is by contrasting cases where T occurs with those where T does not occur in circumstances that are the same as the first with respect to all other factors affecting O other than the occurrence of T and its downstream effects.

My discussion aims to show that neither of these suppositions is correct.[4] Nor do we take them to be correct, at least if the dictum 'actions speak louder than words' is to be believed. We all regularly, in daily life and in professional practice, bet on causal claims about single individuals and guide our actions by these bets without the aid of comparison. Juries decide whether the defendant committed the crime generally without consulting a case just like this one except for the defendant's actions; I confidently infer that it was my second daughter (not the first, not my granddaughter, not Santa) who slipped *Northanger Abbey* into my Christmas stocking; and the NASA investigating team decided that the failure of an O-ring seal in the right solid rocket booster caused the Challenger disaster in which all seven crew were killed.

It might be objected that these causal judgments are made without the rigor demanded in science and wished for in policy. That would be surprising if it were generally true since we treat a good many of these as if we can be reasonably certain of them. Some 975 days after the Challenger disaster, Space Shuttle Discovery – with redesigned solid rocket boosters – was launched with five crew members aboard (and it returned safely four days later). Though not much of practical importance depends on it, I am sure who gave me *Northanger Abbey*. By contrast, people's lives are seriously affected by the verdicts of judges, juries, and magistrates. Though we know that mistakes here are not uncommon, nobody suggests that our abilities to draw singular causal conclusions in this domain are so bad that we might as well flip a coin to decide on guilt or innocence.

I take it to be clear that singular causal claims like these can be true or false, and that the reasoning and evidence that backs them up can be better or worse. The question I address in Section 2.3, with a 'potted' example in Section 2.4, is: What kinds of information make good evidence for singular causal claims about the results of policy interventions, both post-hoc evaluations – 'Did this intervention achieve the targeted outcome when it was implemented here in this individual case?' – and ex ante predictions – 'Is this intervention likely to produce the targeted outcome if implemented here

[3] Cf. Menzies (2014).

[4] For a more detailed discussion, see Cartwright (2017a).

in this individual case?' I believe that the catalogue of evidence types I outline wears its plausibility on its face. But I do not think that is enough. Plausible is, ceteris paribus, better than implausible, but it is better still when the proposals are grounded in theory – credible, well argued, well-warranted theory. To do this job I turn to a familiar theory that is commonly used to defend other conventional scientific methods for causal inference, from randomized controlled trials (RCTs) to qualitative comparative analysis, causal Bayes nets (Bayesian networks) methods, econometric instrumental variables, and others. In Section 2.5, I outline this theory and explain how it can be used to show that the kinds of facts described in the evidence catalogue *are* evidence for causation in the single case.

So, what kinds of facts should we look for in a case study to provide evidence about a singular casual claim there – for instance, a claim of the kind we need for program evaluation: Did this program/treatment (T) as it was implemented in this situation (S) produce an outcome (O) of interest here? Did T cause O in S?

I call the kinds of evidence one gets from case studies for singular causal claims *individualized* evidence. This is by contrast with RCTs, which provide what I call *anonymous* evidence for singular causal claims. I shall explain this difference before proceeding to my catalogue because it helps elucidate the relative advantages and disadvantages of RCTs versus case studies for establishing causal claims.

2.2 What We Can Learn from an RCT

Individualized evidence speaks to causal claims about a particular identified individual; *anonymous* evidence speaks about one or more unidentified individuals. RCTs and group-comparison observational studies provide anonymous evidence about individual cases. This may seem surprising since a standard way of talking makes it sound as if RCTs establish general causal claims – 'It works' – and not claims about individuals at all. But RCTs by themselves establish a claim only about averages, and about averages only in the population enrolled in the experiment. What kind of claim is that? To understand the answer a little formalism is required. [See Appendix 2.1 for more complete development.]

A genuinely positive effect size in an RCT where the overall effects of other 'confounding' variables are genuinely balanced between treatment and control groups – let's call this an 'ideal' RCT – would establish that at least some individuals in the study population were caused by the treatment to have the targeted outcome. This is apparent in the informal argument that positive results imply causal claims: ‘If there are more cases of the outcome in the treatment than in the control group, something must have caused this. If the only difference between the two groups is the treatment and its downstream effects, then the positive outcomes of at least some of the individuals in the treatment group must have been caused by the treatment.’

This is established more formally via the rigorous account of RCT results in common use that traces back to Rubin (1974) and Holland (1986), which calls on the kind of theory appealed to in Section 2.5. We assume that whether one factor causes another in an individual is not arbitrary but that there is something systematic about it. There is a fact of the matter about what factors at what levels in what combinations produce what levels for the outcome in question for each individual. Without serious loss of generality, we can represent all the causal possibilities that are open for an individual *i* in a simple linear equation, called a *potential outcomes equation*:

$$POE(1): \ O(i)c{=}\ \alpha(i)T(i) + W(i)$$

In this equation the variable *O* on the left represents the targeted outcome; *c*= signifies that the two sides of the equation are equal and that the factors on the right are causes of those on the left. *T(i)*, which represents the policy intervention under investigation, may or may not genuinely appear there; that is, *α(i)* may be zero. The equation represents the possible values the outcome can take given various combinations of values a complete set of causes for it takes. *W(i)* represents in one fell swoop all the causes that might affect the level of the outcome for this individual that do not interact with the treatment.[5] *α* represents the overall effect of factors that *interact* with the treatment. ‘Interact’ means that the amount the treatment contributes to the outcome level for individual *i* depends on the value of *α(i)*. Economists and statisticians call these ‘interactive’ variables; psychologists tend to call them ‘moderator’ variables; and philosophers term them ‘support’ variables. For those not familiar with support factors, consider the standard philosopher’s example of striking a match to produce a flame. This only works if there is oxygen present; oxygen is a support factor without which the striking will not produce a flame.

[5] *W(i)* can include a variable that represents a pure individual effect not shared with others in the population.

Interactive/support variables really matter to understanding the connection between the statistical results of an RCT and the causal conclusions inferred from them. The statistical result that is normally recorded in an RCT is the effect size. 'Effect size' can mean a variety of things. But all standard definitions make it a function of this: the difference in outcome means between treatment and control groups. What can this difference in the average value of the outcome in the two groups teach us about the causal effects of the treatment on individuals enrolled in the experiment? What can readily be shown is that in an ideal RCT this difference in means between treatment and control is the mean value of *α(i)*, which represents the support factors – the mean averaged across all the individuals enrolled in the experiment. So the effect size is a function of the mean of the support/interactive variables – those variables that determine whether, and to what extent, the treatment can produce the outcome for the individual. If the average of *α(i)* is not zero, then there must be at least some individuals in that population for which *α(i)* was not zero. That means that for some individuals – though we know not which – *T* genuinely did contribute to the outcome. Thus, we can conclude from a positive mean difference between treatment and control in an ideal RCT that '*T* caused *O* in some members of the population enrolled in the experiment.'[6]

You should also note one other feature of *α(i)*. Suppose that we represent the value of the policy variable in the control group from which it is withheld by *0*. This is another idealization, especially for social experiments and even for many medical ones, where members of the control groups may manage to get the treatment despite being assigned to control. But let's suppose it. Then $\alpha(i)T(i) - \alpha(i)C(i) = \alpha(i)T(i) - 0 = \alpha(i)T(i)$, letting *C* represent the value of the treatment when that treatment is not experienced. So *α(i)* represents also the 'boost' to *O* that *i* gets from receiving the policy treatment. This is often called 'the individual treatment effect'.

When could we expect the same positive average effect size in an RCT on a new population? In the abstract that is easy to say. First, *T* must be *capable* of producing *O* in the new population. There must be possible support factors that can get it to work. If there aren't, no amount of *T* will affect *O* for anyone. Again, philosophers have a potted example: No amount of the fertility drug Clomiphene citrate will make any man get pregnant. In

[6] It may be useful to be reminded that the reverse is not true. The mean in treatment and control groups can be the same not only because the treatment is ineffective but also if it is helpful to some and harmful to others and the effects averaged over the treatment group balance out.

development studies we might use Angus Deaton's (2010) fanciful example of a possible World Bank proposal to reduce poverty in China by building railway stations, a proposal that is doomed to failure when looked at in more detail because the plan is to build them in deserts where nobody lives. Then the two experiments will result in the same effect size just in case the mean of *T*'s support factors is the same in the two. And how would we know this? That takes a great deal of both theoretical and local knowledge about the two populations in question – knowledge that the RCTs themselves go no way toward providing.[7]

Much common talk makes it sound as if RCTs can do more, in particular that they can establish what holds generally or what can be expected in a new case. Perhaps the idea is that if you can establish a causal conclusion then somehow, because it is causal, it is general. That's not true, neither for the causal results established for some identified individuals in an RCT nor for a causal result for a single individual subject that might be established in a case study. Much causality is extremely local: local to toasters of a particular design, to businesses with a certain structure, to fee-paying schools in university towns in the south of England, to families with a certain ethnic and religious background and immigration history . . . The tendency to generalize seems especially strong if 'the same' results are seen in a few cases – which they seldom are, as can be noted from a survey of meta-analyses and systematic reviews. But that is induction by simple enumeration, which is a notoriously bad way to reason (swan 1 is white, swan 2 is white . . . so all the swans in Sydney Harbour are white).

A study – no matter whether it is a case study or it uses the methodology of the RCT, Bayes nets methods for causal inference, instrumental variables, or whatever – by itself can only show results about the population on which the data is collected. To go beyond that, we need to know what kinds of results travel, and to where. And to do that takes a tangle of different kinds of studies, theories, conceptual developments, and trial and error. This is underlined by work in science studies[8] and by recent philosophical work on evidence and induction. See, for instance, John Norton's (2021) material theory of induction: Norton argues that inductive inferences are justified by facts, where facts include anything from measurement results to general

[7] For further discussion, see Cartwright and Hardie (2012). For a wonderful technical treatment of conditions under which different results travel from one population to another, see Bareinboim and Pearl (2013).

[8] Cf. Hasok Chang's (2007) *Inventing Temperature* or Peter Howlett and Mary Morgan's (2010) *How Well Do Facts Travel?*

principles. Parallel lessons follow from the theory of evidence I endorse (Cartwright 2013), the *argument theory*, in which a fact becomes evidence for a conclusion in the context of a good argument for that conclusion, an argument that inevitably requires other premises.

What I want to underline here with respect to RCTs is that, without the aid of lots of other premises, their results are confined to the population enrolled in the study; and what a positive result in an ideal RCT shows is that the treatment produced the outcome in some individuals in that population. For all we know these may be the only individuals in the world that the treatment would affect that way. The same is true if we use a case study to establish that T caused O in a specific identified individual. Perhaps this is extremely unlikely. But the study does nothing to show that; to argue it – either way – requires premises from elsewhere.

I also want to underline a number of other facts that I fear are often underplayed.

- The RCT provides *anonymous* evidence. We may be assured that T caused O in some individuals in the study population, but we know not which. I call this 'Where's Wally?' evidence. We know he's there somewhere, but the study does not reveal him.
- The study establishes an average; it does not tell us how the average is made up. Perhaps the policy is harmful as well as beneficial – it harms a number of individuals, though on average the effect is positive.
- We'd like to know about the variance, but that is not so easy to ascertain. Is almost everyone near the average or do the results for individuals vary widely? The mean of the individual effect sizes can be estimated directly from the difference in means between the treatment and the control groups. But the variance cannot be estimated without substantial statistical assumptions about the distribution. Yet one of the advantages of RCTs is supposed to be that we can get results without substantial background assumptions.
- I have been talking about an ideal RCT in a very special sense of 'ideal': one in which the net effect of confounding factors is genuinely balanced between treatment and control. But that is not what random allocation guarantees for confounders even at baseline. What randomization buys is balance 'in the long run'. That means that if we did the experiment indefinitely often on exactly the same population, the observed difference in means between treatment and control groups would converge on the true difference.

- That's one reason we want experiments to have a large number of participants: it makes it more likely that what we observe in a single run is not far off the true average, though we know it still should be expected to be off a bit, and sometimes off a lot. Yet many social experiments, including many development RCTs, are done on small experimental populations.
- Randomization only affects the baseline distribution of confounders. What happens after that? Blinding is supposed to help control differences, but there are two problems. First, a great many social experiments are poorly blinded: often everybody knows who is in treatment versus control – from the study subjects themselves to those who administer the policy to those who measure the outcomes to those who do the statistical analyses – and all of these can make significant differences. Second, without reasonable local background knowledge about the lives of the study participants (be they individuals or villages), it is hard to see how we have reason to suppose that no systematic differences affect the two groups post randomization.
- Sometimes people say they want RCTs because RCTs measure average effect sizes and we need these for cost–benefit analyses. They do, and we do. But the RCT measures the average effect size in the population enrolled in the experiment. Generally, we need to do cost–benefit analysis for a different population, so we need the average effect size there. The RCT does not give us that.

I do not rehearse these facts to attack RCTs. RCTs are a very useful tool for causal inference – for inferring anonymous singular causal claims. I only list these cautions so that they will be kept in mind in deciding which tool – an RCT or a case study or some other method or some combination – will give the most reliable inference to singular causal claims in any particular case.

I turn now to the case study and how it can warrant singular causal claims – in this case, individualized ones.

2.3 A Category Scheme for Types of Evidence for Singular Causation That a Case Study Can Provide

Suppose a program *T* has been introduced into a particular setting S in hopes of producing outcome *O* there. We have good reason to think *O* occurred. Now we want to know whether *T*, as it was in fact implemented in *S*, was (at least partly) responsible.[9] What kinds of information should we try to collect

[9] Material in this section and the next draws on Cartwright (2017a, 2017b).

in our case study to provide evidence about this? In this section I offer a catalogue of types of evidence that can help. I start by drawing some distinctions. However, it is important to make a simple point at the start. I aim to lay out a catalogue of kinds of evidence that, if true, can speak for or against singular causal claims. How compelling that evidence is will depend on:

- how strong the link, if any, is between the evidence and the conclusion,
- how sure we can be about the strength of this link, and
- how warranted we are in taking the evidence claim to be true.

All three of these are hostages to ignorance, which is always the case when we try to draw conclusions from our evidence. In any particular case we may not be all that sure about the other factors that need to be in place to forge a strong link between our evidence claim and our conclusion, we may worry whether what we see as a link really is one, and we may not be all that sure about the evidence claim itself. The elimination of alternatives is a special case where the link is known to be strong: If we have eliminated alternatives then the conclusion follows without the need for any further assumptions. But, as always, we still face the problem of how sure we can be of the evidence claim. Have we really succeeded in eliminating all alternatives? No matter what kind of evidence claim we are dealing with, it is a rare case when we are sure our evidence claims are true and we are sure how strong our links are, or even if they are links at all. That's why, when it comes to evidence, the more the better.

The first distinction that can help provide a useful categorization for types of evidence for singular causal claims is that between direct and indirect evidence:

- ***Direct:*** Evidence that looks at aspects of the putative causal relationship itself to see if it holds.
- ***Indirect:*** Evidence that looks at features outside the putative causal relationship that bear on the existence of this relationship.

Indirect. The prominent kind of indirect evidence is evidence that helps eliminate alternatives. If *O* occurred in *S*, and anything other than *T* has been ruled out as a cause of *O* in *S*'s case, then *T* must have done it. This is what Alexander Bird (2010, 345) calls 'Holmesian inference' because of the famous Holmes remark that when all the other possibilities have been eliminated, what remains must be responsible even if improbable. RCTs provide indirect evidence, eliminating alternative explanations by (in the ideal) distributing all the other possible causes of *O* equally between treatment and control groups. But we don't need a comparison group to do this. We can do this in the case study as well, if we know enough about what the other causes might be like, and/or about the history of the situation *S*. We do this in physics experiments regularly. But we don't need physics to do it. It is, for instance, how I know it was my cat that stole the pork chop from the frying pan while I wasn't looking.

Direct. I have identified at least four different kinds of direct evidence possible for the individualized singular causal claim that *T* caused *O* in *S*:

1. The character of the effect: Does *O* occur at the time, in the manner, and of the size to be expected had *T* caused it? (For those who are familiar with his famous paper on symptoms of causality, Bradford Hill (1965) endorses this type of evidence.)
2. Symptoms of causation: Not symptoms that *T* occurred but symptoms that *T* caused the outcome, side effects that could be expected had *T* operated to produce *O*. This kind of inference is becoming increasingly familiar as people become more and more skilled at drawing inferences from 'big data'. As Suzy Moat puts it, "People leave this large amount of data behind as a by-product of simply carrying on with their lives." Clever users of big data can reconstruct a great deal about our individual lives from the patterns they find there.[10]
3. Presence of requisite support factors (moderator/interactive variables): Was everything in place that needed to be in order for *T* to produce *O*?
4. Presence of expectable intermediate steps (mediator variables): Were the right kinds of intermediate stages present?

Which of these types of evidence will be possible to obtain in a given case will vary from case to case. Any of them that we can gather will be equally relevant for post-hoc evaluation and for ex ante prediction, though we certainly won't ever be able to get evidence of type 2 before the fact. I am currently engaged in an NSF-funded research project, *Policy Prediction: Making the Most of the Evidence*, that aims to use the situation-specific causal equations model (SCEM) framework sketched in Section 2.5 to expand this catalogue of evidence types and to explore more ways to use it for policy prediction.

[10] At a *Spaces of Evidence* conference, Goldsmiths, University of London, Sept. 26, 2014. See Moat et al. (2014).

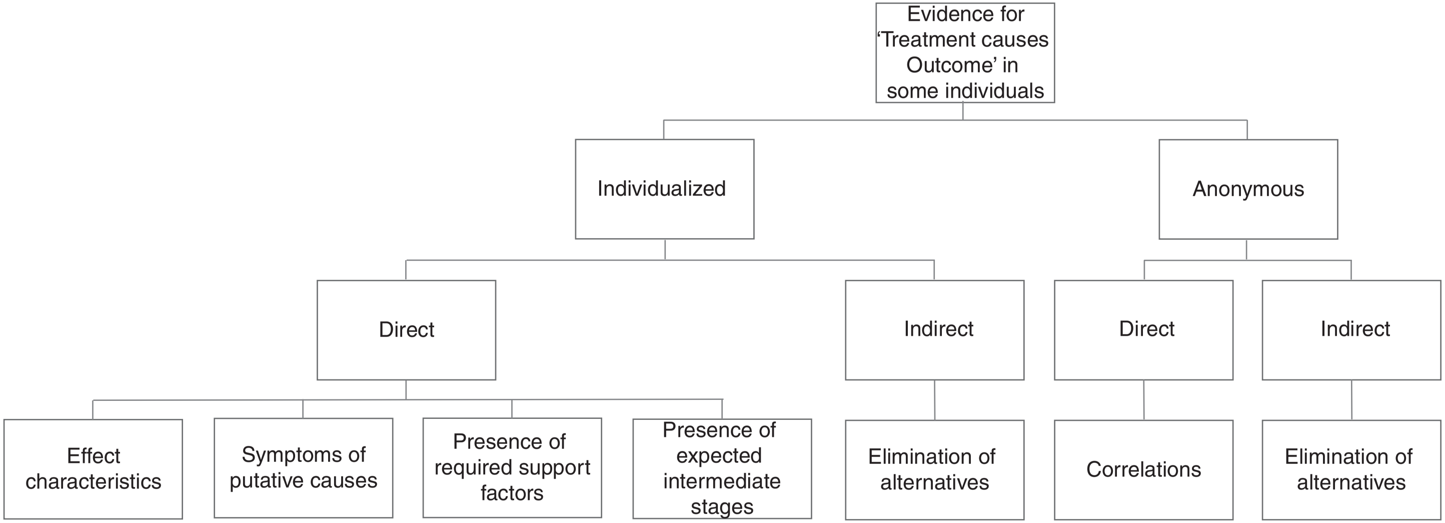
Evidence for 'Treatment causes Outcome' in some individuals
Individualized
Anonymous
Direct
Indirect
Direct
Indirect
Effect characteristics
Symptoms of putative causes
Presence of required support factors
Presence of expected intermediate stages
Elimination of alternatives
Correlations
Elimination of alternatives

Figure 2.1 Categories of evidence

2.4 A Diagrammatic Example

Let me illustrate with one of those diagrammatic examples we philosophers like, this one constructed from my simple-minded account of how an emetic works. It may be a parody of a real case study, but it provides a clear illustration of each of these types of evidence.

Imagine that yesterday I inadvertently consumed a very harmful poison. Luckily, I realized I had done so and thereafter swallowed a strong emetic. I vomited violently and have subsequently not suffered any serious symptoms of poisoning. I praise the emetic: It saved me! What evidence could your case study collect for that?

- Elimination of alternatives: There are very low survival rates with this poison. So it is not likely my survival was spontaneous. And there's nothing special about me that would otherwise explain my survival having consumed the poison. I don't have an exceptional body mass, I hadn't been getting slowly acclimatised to this poison by earlier smaller doses, I did not take an antidote, etc.
- Presence of required support factors (other factors without which the cause could not be expected to produce this effect): The emetic was swallowed before too much poison was absorbed from the stomach.
- Presence of necessary intermediate step: I vomited.
- Presence of symptoms of the putative causes acting to produce the effect: There was much poison in the vomit, which is a clear side effect of the emetic's being responsible for my survival.
- Characteristics of the effect: The amount of poison in the vomit was measured and compared with the amount I had consumed. I suffered just the effects of remaining amount of poison; and the timing of the effect and size were just right.

2.5 Showing This Kind of Information Does Indeed Provide Evidence about Singular Causation

I developed the scheme in Section 2.3 for warranting singular causal claims bottom-up by surveying case studies in engineering, applied science, policy evaluation, and fault diagnoses, inter alia. But a more rigorous grounding is possible: these types all provide information relevant for filling in features of a *situation-specific causal equations model* (SCEM). Once you see what a SCEM is, this is apparent by inspection, so I will not belabor that point. Instead, I will spend time defending the SCEM framework itself.

A SCEM is a set of equations that express (one version of) what is sometimes called the 'logic model' of the policy intervention: a model of how the policy treatment T is supposed to bring about the targeted outcome O, step by step. Each of the equations is itself what in Section 2.2 was called a 'potential outcomes equation'. (In situations where the kind of quantitative precision suggested by these equations seems impossible or inappropriate, there is an analogous Boolean form for yes–no variables, familiar in philosophy from Mackie (1965) and in social science from qualitative comparative analysis [e.g., Rihoux and Ragin 2008].)

To build a SCEM, start with the outcome *O* of interest. Just what should the policy have led to at the previous stage that will produce *O* at the final stage? Let's call that 'O_{-1}'. Recalling that a single cause is seldom enough to produce an effect on its own, what are the support factors necessary for O_{-1} to produce *O*? Represent the net effect of all the support factors by 'α_{-1}'. Establishing that these support factors were/will be in place or not provides important evidence about whether *O* can be brought about by O_{-1}. If not, then certainly *T* cannot produce *O* (at least not in the way you expect). Consider as well what other factors will be in place at the penultimate stage that will affect *O*. These affect the size or level of *O*. You want to know about those because they provide alternative explanations for the level of *O* that occurs; they are also relevant for judging the size *T*'s contribution would have to be if *T* were to contribute to the outcome. Represent the net effect of all these together by 'W_{-1}'. How *O* depends on all these factors can then be represented in a potential outcomes equation like this:

$$POE\ (2):\ \ O(i)\ c{=}\ \alpha_{-1}(i)O_{-1}(i) + W_{-1}(i).$$

Work backwards, step by step, constructing a potential outcomes equation for each stage until the start, where *T* is introduced. The resulting set of equations is the core of the SCEM for this case.

But there is more. Think about the support factors (represented by the *α*s) that need to be in place at each stage. These are themselves effects; they have a causal history that can be expressed in a set of potential outcomes equations that can be added to the core SCEM. This is important information too: Knowing about the causes of the causes of an effect is a clue to whether the causes will occur and thus to whether the effect can be expected. The factors that do not interact with O_{-1} (represented by W_{-1}) but that also affect *O* have causal histories as well that can be represented in a series of potential outcomes equations and added to the SCEM. So too with all the *W*s in the chain. For purposes of evaluation, we may also want to include equations in which *O* figures as a cause since seeing that the effects of *O* obtain gives good evidence that *O* itself occurred. We can include as much or as little of the causal histories of various variables in the SCEM as we find useful.

I am not suggesting that we can construct SCEMs that are very complete, but I do suggest that this is what Nature does. Even in the single case, what causes what is not arbitrary – at least not if there is to be any hope that we can make reasonable predictions, explanations, and evaluations. There is a system to how Nature operates, and we have learned that generally this is what the system is like: Some factors *can* affect O in this individual and some *cannot*. All those that can affect an outcome appear in Nature's own potential outcomes equation for that outcome. Single factors seldom contribute on their own so the separate terms in Nature's equations will generally consist of combinations of mutually interacting factors. So Nature's equations look much like ours. Or, rather, when we do it well, ours look much like Nature's since hers are what we aim to replicate.

So: A successful SCEM for a specific individual provides a concise representation of what causal sequences are possible for that individual given the facts about that individual and its situation – what values the quantities represented can take in relation to values of their causes and effects. Some of the features represented in the SCEM will be ones we can influence, and some of these are ones we would influence in implementing the policy; others will take the values that naturally evolve from their causal past. The interpretation of these equations will become clearer as I defend their use.

I offer three different arguments to support my claim that SCEMs are good for treating singular causation: 1) their use for this purpose is well developed in the philosophy literature; 2) singular causation thus treated satisfies a number of common assumptions; 3) the potential outcomes equations that make up a SCEM are central to the formal defense I described in Section 2.2 that RCTs can establish causal conclusions.[11]

1) The SCEM framework is an adaptation for variables with more than two values of J. L. Mackie's (1965) famous account in which causes are INUS conditions for their effects. In the adaptation, causes are INUS conditions

[11] As mentioned in Section 2.2, they are similarly central to the defense of a variety of other methods for causal inference, though I do not show that here.

for *contributions* to the effect,[12] where an INUS condition for a contribution to O(i) is an **I**nsufficient but **N**ecessary part of an **U**nnecessary but **S**ufficient condition for a contribution to it. Each of the additive terms ($\alpha(i)T(i)$ and $W(i)$) on the right of the equation $O(i)\ c= \alpha(i)T(i) + W(i)$ represents a set of conditions that together are *sufficient* for a contribution to $O(i)$ but they are *unnecessary* since many things can contribute to O; and each component of an additive term (e.g., $\alpha(i)$ and $T(i)$) is an *insufficient* but *necessary* part of it – both are needed and neither is enough alone. This kind of situation-specific causal equations model for treating singular causation is also familiar in the contemporary philosophy of science literature, especially because of the widely respected work of Christopher Hitchcock.[13]

2) The SCEM implies a number of characteristics for singular causal relations that they are widely assumed to have:
   - the causal relation is irreflexive (nothing causes itself)
   - the causal relation is asymmetric (if T causes O, O does not cause T)
   - causes occur temporally before their effects
   - there are causes to fix every effect
   - causes of causes of an effect are themselves causes of that effect (since substituting earlier causes of the causes in an equation yields a POE valid for a different coarse-graining of the time)[14]
   - causal relations give rise to noncausal correlations.[15]
3) Each equation in a SCEM is a potential outcomes equation of the kind that is used in the Rubin/Holland argument I laid out in Section 2.2 to show that RCTs can produce causal conclusions: A SCEM is simply a reiteration of the POE used to represent singular causation in the treatment of RCTs, expanded to include causes of causes of the targeted outcome and, sometimes, further effects as well. So, if we buy the Rubin/Holland argument about why a positive difference in means between treatment and control groups provides evidence that the treatment has caused the outcome in at least some members of the treatment group, it seems we are committed to taking POEs, and thus SCEMs, as a good

[12] Note: stating that all causes are INUS conditions does not imply that all INUS conditions are causes.

[13] Cf. Hitchcock (2007).

[14] Philosophers sometimes reject this assumption, but it is important for predicting effects separated by longish time periods from the policy initiation.

[15] For example, consider a cause c with two effects, e_1 and e_2, with no other causes. Supposing determinism, e_1 obtains if and only if e_2 obtains. That is not among the causal equations. But it obtains on account of them.

representation of the causal possibilities open to individuals in the study population.

Warning: Equations like these are sometimes treated as if they represent 'general causal principles'. That is a mistake. To see why, it is useful to think in terms of a threefold distinction among equations we use in science and policy, and similarly for more qualitative principles:

- Equations and principles that represent the context-relative causal possibilities that obtain for a specific single individual, as in the SCEMs discussed here.
- Equations and principles that represent the context-relative causal possibilities for a specific population. These often look just like a SCEM so it appears as if the causal possibilities are the same for every member of the population. This can be misleading for two reasons. First, for some individuals in the population some of the $\alpha(i)$s may be fixed at *0* so that the associated cause can never contribute to the outcome for them. Second, the $W(i)$s can contain a variable that applies only to the single individual i (as noted in footnote 5). So there can be unique causal possibilities for each member of the population despite the fact that the equation makes it look as if they are all the same.
- Equations and principles that hold widely. I suggest reserving the term 'general principles' for these, which are relatively context free, like the law of the lever or perhaps 'People act so as to maximize their expected utility.' These are the kinds of principles that we suppose ground the single-case causal possibilities represented in SCEMs and the context-relative principles that describe the causal possibilities for specific populations. These general principles tend to employ very abstract concepts, by contrast with the far more concrete, operationalizable ones that describe study results on individuals or populations – abstract concepts such as 'utility', 'force', 'democracy'. They are also generally different in form from SCEMs. Think, for instance, about the form of Maxwell's equations, which ground the causal possibilities for any electromagnetic device: these are not SCEM-like in form at all. It is in an instantiation of these in a real concrete arrangement located in space and time that genuine causal possibilities, of the kind represented in SCEMs, arise.

I note the differences between equations representing general principles and those representing causal possibilities for a single case or for a specific population to underline that knowing general principles is not enough to tell us what we need to know to predict policy outcomes for specific individuals, whether these are individual students or classes or villages, considered as a whole, or specific populations in specific places. Knowing Maxwell's principles will not tell you how to repair your Christmas-tree lights. For that you need context-specific local knowledge about what the local arrangements are that call different general principles into play, both together and in sequence. That's what will enable you to build a good SCEM that you can use for predicting and explaining outcomes. The same unfortunately is true for the use of general principles to predict the results of development and other social policies. Good general principles should be very reliable, but it takes a lot of thinking and a lot of local knowledge to figure out how to deploy them to model concrete situations. This is one of the principal reasons why we need case studies.

Thinking about how local arrangements call different general principles into play or not is key to how to make good use of our general knowledge to build local SCEMs. Consider a potted version of the case of the failure of the class-size reduction program that California implemented in 1996/97 based on the successes of Tennessee's STAR project (which was attested by a good RCT) and Wisconsin's SAGE program. Let us suppose for purposes of illustration that these three general principles obtain widely:

- Smaller classes are conducive to better learning outcomes.
- Poor teaching inhibits learning.
- Poor classroom facilities inhibit learning.

Imagine that in Tennessee there were good teacher-training schools with good routes into local teaching positions and a number of new schools with surplus well-equipped classrooms that had resulted from a vigorous, well-funded school-building program. In California there was a great deal of political pressure and financial incentivization to introduce the program all at once (it was rolled out in most districts within three months of the legislation being passed); there were few well-trained unemployed teachers and no vigorous program for quick recruitment; and classrooms, we can suppose, were already overcrowded. These arrangements in California called all three principles into play at once; thus – so this story goes – the good effects promised by the operation of the first principle were outweighed by the harmful effects of the other two. Learning outcomes did not improve across the state, and in some places they got worse.[16] The arrangements in

[16] For a serious account of what happened, see Stecher and Bohrnstedt (2002).

Tennessee called into play only the first principle, which accounts for the improved outcomes there.

How would you know whether to expect the results in California to match those of Tennessee and Wisconsin? Not by looking for superficial 'similarities' between the two. I recommend a case study, one that builds a SCEM for California, modelling the sequential steps by which the policy is supposed to achieve the targeted outcomes and then modelling what factors are needed in order for each step to lead to the next and what further causes are supposed to ensure that these factors are in place. We can't do this completely, but reviewing the California case, it seems there was ample evidence – evidence of the kinds laid out in the catalogue of Section 2.3 – to fill in enough of the SCEM to see that a happy outcome was not to be expected.

2.6 Conclusion

How much evidence of the kinds in my catalogue and in what combinations must a case study deliver, and how secure must it be, in order to provide a reasonable degree of certainty about a causal claim about the case? There's no definitive answer. That's a shame. But this is not peculiar to case studies; it is true for all methods for causal inference.

Consider the RCT. If we suppose the treatment does satisfy the independence assumptions noted in Appendix 2.1, we can calculate how likely a given positive difference in means is if the treatment had no effect and the difference was due entirely to chance. But for most social policy RCTs there are good reasons to suppose the treatment does not satisfy the independence assumptions. The allocation mechanism often is not by a random-outcome device; there is not even single blinding let alone the quadruple we would hope for (of the subjects, the program administrators and overseers, those who measure outcomes, and those who do the statistical analysis); numbers enrolled in the experiment are often small; dropouts, noncompliance, and control group members accessing the treatment outside the experiment are not carefully monitored; sources of systematic differences between treatment and control groups after randomization are not well thought through and controlled; etc. – the list is long and well known. Often this is the best we can do, and often it is better than nothing. The point is that there are no formulae for how to weigh all this up to calculate what level of certainty the experiment provides that the treatment caused the outcome in some individuals in the experimental population. Similarly with all other methods of causal inference. Some things can be calculated – subject to assumptions. But there is seldom a method for calculating how the evidence that the assumptions are satisfied stacks up, and we often have little general idea about what that evidence should even look like. Judgment – judgment without rules to fall back on – is required in all these cases. I see no good arguments that the judgments are systematically more problematic in case studies than anywhere else.

The same holds when it comes to expecting the same results elsewhere. Maybe if you have a big effect size in an RCT with lots of subjects enrolled and good reason to think that the independence assumptions were satisfied, you have reason to think that in a good number of individuals the treatment produced the outcome. For a single case study, you can have at best good reason to think that the treatment caused the outcome in one individual. Perhaps knowing it worked for a number of individuals gives better grounds for expecting it to work in the next. Perhaps not. Consider economist Angus Deaton's (2015) suggestions about St. Mary's school, which is thinking about adopting a new training program because a perfect RCT elsewhere has shown it improves test scores by X. But St. Joseph's down the road adopted the program and got Z. What should St. Mary's do? It is not obvious, or clear, that St. Joseph's is not a better guide than the RCT, or indeed an anecdote about another school. After all, St. Mary's is not the mean, and may be a long way from it. Which is a better guide – or any guide at all – depends on how similar, in just the right ways, the individual/individuals in the study are to the new one we want predictions about. And how do we know what the right ways are? Well, a good case study at St. Joseph's can at least show us what mattered for it to work there, which can be some indication of what it might take to work at St. Mary's since they share much underlying structure.[17] In this case it looks like the advantage for exporting the study result may lie with the case study and not with the higher numbers.

Group-comparison studies do have the advantage that they can estimate an effect size – for the study population. That may be just what we need – for instance, in a post-hoc evaluation where the program contractors are to be paid by size of result. But we should beware of the assumption that this number is useful elsewhere. We have seen that it depends on the mean value

[17] Here's yet another source of uncertainty in both cases. The – often unknown or ill-understood – underlying structure matters to what can help a cause to operate. What enables a cause to work in one given underlying structure need not enable it to work where other structures obtain. Putting gas in my Volvo enables the car to go when I turn the ignition on, but not in a diesel Audi 3; and reducing class sizes in Tennessee and Wisconsin improved learning outcomes, but not in California.

of the net contribution of the interactive/support factors in the study population. It takes a lot of knowledge to warrant the assumption that the support factors at work in a new situation will have the same mean.

What can we conclude in general, then, about how secure causal conclusions from case studies are or how well they can be exported? Nothing. But other methods fare no better.

There is one positive lesson we can draw. We often hear the claim that case studies may be good for suggesting causal hypotheses but it takes other methods to test them. That is false. Case studies can test causal conclusions. And a well-done case study can establish causal results more securely than other methods if they are not well carried out or we if have little reason to accept the assumptions it takes to justify causal inference from their results.

Appendix 2.1

The Rubin/Holland analysis, which is also widely adopted by economists discussing RCTs, begins with a singular counterfactual difference: that between the value that the outcome (say $x_k(i)$) would have in the individual case i were i subject to the treatment ($x_k{}^T(i)$) and the value it would have in i were i subject to the control ($x_k{}^C(i)$). It is assumed that the possible values x_k can take for i are determined[18] by a complete set of possible causes of x_k that might act on i during the relevant time period given the actual situation of i, including possibly the treatment T (which in this simple case gets value 1 in the treatment group and 0 in the control). This gets represented in the potential outcomes equation:

POE: $x_k(i)\ c\!= \alpha_T(i)T(i) + \Sigma\alpha_j(i)x_j(i)$

In this equation the variables on the left represent the targeted outcome; $c\!=$ signifies that the two sides of the equation are equal and that the factors on the right are causes of those on the left. $T(i)$ may or may not genuinely appear there (i.e., $\alpha_T(i)$ may be zero). The equation represents the possible value the outcome can take given various combinations of values a complete set of causes for it take. Besides the treatment there are J possible additive causes as well as those that make up the interactive factor $\alpha_T(i)$ (which may turn out to be 0), most unknown or unobservable.

[18] The scheme can be adapted to deal with merely probabilistic causation, but I won't do that here to keep notation simple.

Now consider treatment (T = 1) and control (T = 0) groups and calculate averages. Imagine that random assignment and blinding have succeeded as hoped in ensuring that T is orthogonal in the mean to the net effect of other causal factors (*$\alpha_T(i)$* and *$\Sigma\alpha_j(i)x_j(i)$*), in which case, using *Exp* for 'expectation'

$$Exp\,(x_k^T(i) - x_k^C(i)) = Exp\,x_k^T(i) - Exp\,x_k^C(i) = Exp\alpha_T(i) \qquad (1)$$

So the middle term (the difference in means between the treatment and control groups, which is observable) is an unbiased estimator of *$Exp\ \alpha_T(i)$*. Given the causal interpretation proposed of the potential outcomes equation, a genuinely positive effect size shows that *$\alpha_T(i) \neq 0$* for some *i*: that is, that (in the long run of experiment repetitions) the treatment will have caused the outcome in at least some individuals in the treatment group. Note that the observed effect size is the estimated mean of the *individual* treatment effects, which in turn is the estimated mean of *$\alpha_T(i)$*. By inspection *$\alpha_T(i)$* represents the net effect of the interactive/support factors that fix whether, and to what degree, *T* can contribute to x_k in individual case *i*.

Appendix 2.2

A SCEM is a set of equations in block triangular form:[19]

$$x_1(i)\ c{=}\ \mu \qquad (1)$$

$$x_2(i)\ c{=}\ a_{21}(i)\,x_1(i)$$

$$x_n c = \ \sum a_{nj}(i)x_j(i)$$

SCEMs provide a concise representation of what causal sequences are possible in a specific case given the facts about that case – that is, what values the quantities represented can take in relation to values of their causes and effects. Each equation is itself a potential outcomes equation (POE). The variables are time ordered so that for $x_{j<k}$, x_j occurs simultaneous with, or earlier than, x_k. As with a single POE, variables on the left represent effects, one of which will be the targeted outcome; *c=* signifies that the two sides of the equation are equal and that the factors on the right are causes of those on the left.

[19] *Block* triangularity allows for multiple simultaneous effects with the same causes.

Warning. In addition to the warning in Section 2.5, I offer some further, more technical cautions here. The linear simultaneous equations forms that appear in a SCEM are also familiar within social science, for example from the work of Herbert Simon (1957) on causality, in path analysis, in econometrics, as the basis for Judea Pearl's causal Bayesian networks,[20] etc. I say 'warning' because I see two related problems cropping up. First, the equations show relations between quantities but they do not express which populations of individual cases these relations hold for, and often this is not made clear in social science uses. I use these equations for *identified* individual cases. Second, generally some of the variables are labelled 'exogenous' (determined outside the system of equations, indicated by the μ in the first equation) and a joint probability is supposed for them. This supposes some population of individual cases to which this probability applies, but again, which population that is – or why we should suppose there is a probability to be had for those variables in that population – is usually not specified.

References

Bareinboim, E. and Pearl, J. (2013) "A general algorithm for deciding transportability of experimental results," *Journal of Causal Inference*, 1(1), 107–134.

Bird, A. (2010) "Eliminative abduction: Examples from medicine," *Studies in History and Philosophy of Science*, 41(4), 345–352.

Bradford Hill, A. (1965) "The environment and disease: Association or causation?" *Proceedings of the Royal Society of Medicine*, 58(5), 295–300.

Byrne, D. and Ragin, C. (2009) *The SAGE handbook of case-based methods*. Thousand Oaks, CA: Sage.

Cartwright, N. (2013) "Evidence, argument and prediction" in Karakostas, V. and Dieks, D. (eds.) *EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings 2*, pp. 3–18. Also in *Evidence: For Policy and Wheresoever Rigor is a Must*. Order Project Discussion Paper Series. London: London School of Economics. Available at https://nyudri.wordpress.com/initiatives/deaton-v-banerjee/ (accessed December 20, 2021).

Cartwright, N. (2017a) "Single case causes: What is evidence and why" in Chao, H., Chen, S., and Reiss, J. (eds.) *Philosophy of science in practice*. Dordrecht: Springer, pp. 11–24.

Cartwright, N. (2017b) "How to learn about causes in the single case." Durham University: CHESS Working Paper No. 2017–04.

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

[20] Cf. Pearl (2000).

Chang, H. (2007) *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Deaton, A. (2010) "Instruments, randomization and learning about development," *Journal of Economic Literature*, 48(2), 424–455.

Deaton, A. (2015) "Deaton v Banerjee." NYU Development Research Institute. Available at https://nyudri.wordpress.com/initiatives/deaton-v-banerjee/ (accessed January 20, 2020).

Hitchcock, C. (2007) "Prevention, preemption, and the principle of sufficient reason," *Philosophical Review*, 116(4), 495–532.

Holland, P. (1986) "Statistics and causal inference," *Journal of the American Statistical Association*, 81(396), 945–960.

Howlett, P. and Morgan, M. (eds.) (2010) *How well do facts travel?* Cambridge: Cambridge University Press.

Mackie, J. L. (1965) "Causes and conditions," *American Philosophical Quarterly*, 2(4), 245–264.

Moat, H. S., Preis, T., Olivola, C.Y., Liu, C., and Chater, N. (2014) "Using big data to predict collective behavior in the real world," *Behavioral and Brain Sciences*, 37(1), 92–93.

Morgan, M. (2014) "Case studies" in Cartwright, N. and Montuschi, E. (eds.), *Philosophy of social science: A new introduction*. New York: Oxford University Press, pp. 288–307.

Menzies, P. (2014) "Counterfactual theories of causation" in Zalta, E. N. (ed.), *The Stanford encyclopedia of philosophy* (Spring 2014 Edition). Available at: http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/ (accessed January 20, 2020).

Norton, J. (2021) *The material theory of induction*. Calgary: University of Calgary Press.

Pearl, J. (2000) *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rihoux, B. and Ragin, C. (eds.) (2008) *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*. Los Angeles: Sage Publications.

Rubin, D. (1974) "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66(5), 688–701.

Simon, H. (1957) *Models of man: Social and rational*. New York: John Wiley and Sons.

Stecher, B. and Bohrnstedt, G. (2002) "Class size reduction in California: Findings from 1999–00 and 2001–02" [online]. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.694.7009&rep=rep1&type=pdf (accessed January 20, 2020).