To appear in *Philosophical Issues in Psychiatry VI,* Edited by Kenneth Kendler, Peter Zachar, & Josef Parnas. (Cambridge University Press)

Reflections on Some Strategies for Causal Inference in Psychiatry[*]

James Woodward

1. Introduction.

This paper discusses some widely used strategies (not just in psychiatry but elsewhere) for inferring causal relations-- including randomized controlled trials (RCTs) and instrumental variables. In psychiatry, these seem to be most useful in identifying environmental factors that may play a causal role in mental illness or drugs or therapies that may be used in treating mental illness; it is less obvious (at least to me) how these techniques might be used in any very direct way to make inferences about the causal role of genetic factors or neural functioning in mental illness. Hence my focus will be on the use of these strategies in identifying environmental causes. I will discuss both the advantages and limitations of the strategies.

2. Causation in General

Necessity. A common feature of many accounts is that causation should be understood in terms of some mixture of necessity and sufficiency. In modern treatments, "necessity" is generally understood in terms of the idea that a cause is something that, under the right conditions, makes a difference to whether an effect occurs, so that there is a relation of dependence between the cause and the effect. The qualifying phrase "under the right conditions" is required because, for example, *C1* might cause *E* but the difference-making role of *C1* may be masked by the presence of another cause, *C2* of *E.* In such cases, the difference-making role of *C1* may be recovered by removing or inactivating *C2,* thus revealing the dependence of *E* on *C1*, as in the example involving the discovery of path-specific effects discussed below. The idea that causes are difference-makers in this sense is highlighted in "counterfactual" accounts of causation, including the potential outcomes framework associated with Rubin and prominent in statistics and econometrics as well as the interventionist account described below.

Sufficiency. The sufficiency aspect of causation is more subtle. Both in ordinary life and in the biomedical and behavioral sciences, virtually none of the variables we describe as causes are literally sufficient for their effects in the sense that the effects always follow when the causes are present. Smoking causes lung cancer but not always and even when it does, other conditions must be present-- e.g., failure of DNA repair or tumor suppression mechanisms. I suggest below that a natural way of capturing what is right about the sufficiency idea is that causal relationships (or at least the sorts of causal relationships we value and would like to discover) are expected to be at least somewhat *stable* or *invariant,* in the sense that we expect that they will continue to operate across a range of different circumstances. Thus smoking does not always cause lung cancer but it does for different demographic groups, environmental conditions and so on.

[*] Many thanks to Ken Kendler for very helpful comments on an earlier draft.

Although the necessity aspect of causation is well captured by counterfactual and interventionist accounts of the sort described below, I have come to believe (in agreement with recent criticisms from others-- e.g. Deaton and Cartwright, 2018) that such accounts tend to somewhat underweigh the importance of the sufficiency or stability aspect of causation. I will explore some of the consequences of this for causal inference below, arguing that a number of standard inferential techniques also underweigh the sufficiency aspect.

3. Interventionism

In what follows, I assume a broadly interventionist account of causation. Because I have discussed this account elsewhere (e.g. Woodward, 2003) and in order to save space, I will describe just the bare bones. The basic idea is that if you were to perform the right kind of manipulation of *C* (an "intervention") and *E* changes, then *C* causes *E*. "Right kind" means that the manipulation of *C* should be such that any change in *E* resulting from this manipulation should occur "through" *C* and not in some other way. Put more simply, the manipulation should be unconfounded. At a population level, when the target causal notion is the estimation of an average causal effect, randomized control experiments are one widely recognized way of implementing interventions. However, many -- arguably most-- experimental manipulations in sciences like physics, chemistry and even biology that establish the existence of causal effects do *not* involve randomization, although they do make use of interventions in the sense characterized above. When Michael Faraday established that moving a conductor through a magnetic field induced a current in the conductor, he did not employ randomization. The most famous experiments in molecular biology do not make use of randomization. As discussed below, randomization is a rational response to a very particular set of inferential problems that can arise in portions of the biomedical and behavioral sciences -- problems that arise less often in other areas of science.

4. Distinctions Among Causal Concepts.

So far I have been talking about "causation" in a rather generic way. But one of the attractions of interventionist approaches to causation is that they allow for a more fine-grained characterization of a number of distinct causal notions, corresponding to different questions in which we may be interested.

Total or Net Effects. This has to do with overall total causal impact of a change in one variable on another.

Causal Contribution Along a Path. An important contrasting notion is the notion a causal contribution along a path. Consider the following causal structure:

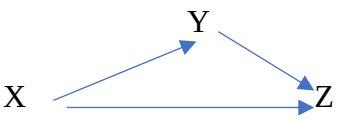


Figure 1

and a set of corresponding structural equations.

$Z = aX + bY$

$Y = cX$

The total effect of *X* on *Z* is the sum of the effects of *X* along the two paths, one direct and one going through *Y* -- hence given by the coefficient $a+bc$. The causal contribution of *X* to *Z* along the direct path is captured by the coefficient *a* and the contribution of *X* to *Z* along the indirect path is given by *bc*. Breaking total causal effects into contributions made along distinct paths (rather than just representing the total effect) is often desirable, perhaps particularly in the biomedical sciences and psychiatry. For example, a gene *G* might contribute to depression *D* via an "inside the skin" pathway directly affecting brain chemistry but also via distinct, outside the skin pathway in which *G* contributes to behavior *B* which creates stressful situations *S* which also contribute to depression. I take it that part of what is involved in the discovery of mechanisms is often the decomposition of total effects into effects along distinct pathways, with distinct mediating variables.

In many (but not all) cases, we can capture this notion of a cause making a contribution to an effect along a path by considering *combinations* of interventions, rather than single interventions as in the characterization of a total effect. In the above example, if we were to intervene to fix the value of one of the mediating variables *B* or *S* at the off position (e.g., by teaching a patient to avoid behavior that causes stressful situations) and, independently of this, intervene to change *G* or its downstream but inside the skin effects, this would (if the postulated causal structure is correct) lead to a change in *D*, thus revealing the role of the inside the skin causal path from *G* to *D* . Similarly, if we could somehow intervene to interfere with the inside the skin path from *G* to *D* (while leaving the outside the skin path through *B* unaffected) while changing *G*, this would reveal the existence of the outside the skin path. As this example brings out, one (of several reasons) why the decomposition of total causal relations into path-specific causal relationships is important is that distinct paths can offer distinct points or variables for intervention and sometimes it may be easier (or more beneficial) to intervene on these than on a total cause. In the example above, an intervention that takes the form of a change in variable *B* may ameliorate the effects of *G* even if we cannot intervene on *G* itself. In psychiatry and psychology, "mediation analysis" is sometimes employed in an effort to achieve this sort of decomposition into paths when one has only observational and not intervention- based information concerning the relation between a candidate mediator and other variables. I discuss this briefly below, noting that it requires very strong assumptions to be reliable.

Average Causal Effect. A third notion of causation (or at least of causal effect) that deserves mention is the notion of average causal effect (ACE) or average treatment effect (ATE). Here an intervention fixes a value for a putative cause for some individuals in a population (the treatment group) and withholds the treatment from other individuals (the control group). If allocation of treatment is randomized, (and other well-known conditions such as the stable unit treatment

value assumption (SUTVA)[1] are satisfied) any "statistically significant" difference in average value for an effect variable between the treatment and control group is taken to support an inference to the average causal effect of the treatment. ACE qualifies as a notion of causation that belongs in the interventionist family because if the RCT is reliable it tells us how an intervention that changes the average value of the treatment variable will change the expected value of the effect variable for the population and circumstances which the RCT represents.

5. Stability.

The interventionist conditions for the various causal notionsjust described can be regarded as *minimal* conditions for causation -- that is, they are conditions that must be satisfied if a relationship is to count as causal at all, as opposed to being a relationship of non-causal association. However, there are further distinctions we may draw among causal relations. I will focus on two of these-- here the aforementioned *stability* (or invariance) and, below, *specificity*. These are among the "aspects" of causation that Bradford Hill (1965) discusses, so here I will be picking up on some of Hill's ideas about causal inference.

Suppose that a relationship satisfies the minimal condition for causation in the sense that it correctly describes how a variable or its expected value will respond to interventions on the individuals within a population in some circumstances, The stability of that causal relationship has to do with the extent to which it continues to hold in other circumstances or populations-- whether it generalizes or "ports". As noted above, the smoking --> lung cancer causal relation is stable in the sense that it holds across many different circumstances and populations-- among different demographic groups, people with different diets and life circumstances and so on. (Note that this is not the same thing as the requirement that relationship be “strong" in the sense that, say, the correlation between smoking and lung cancer be "large"-- the correlation might be modest but hold cross a variety of circumstances.) As I understand stability, it is not an all or nothing matter but rather one of degree: a causal relationship can be more or less stable and stable across some changes in circumstances but not others. We can make a causal claim more precise by specifying this additional information.

Stability is related to what Hill called consistency:

> Has [the relationship] been repeatedly observed by different persons, in different places, circumstances and times?

However, my notion of stability is somewhat different from and stronger than this. It has to do not just with whether an *association* generalizes to different circumstances but whether a *causal relationship*, holding in some local circumstances, also holds in others. It isn't just that smoking and lung cancer are associated in many different circumstances; in addition smoking *causes* lung cancer in many different circumstances. It is entirely possible for an association to hold in many different circumstances but not be causal. This can happen when the association is the

[1] SUTVA requires that the response of each unit in the population should not depend on the treatment assigned to other units. This requirement is violated when there is interference or "spillover" between units, as when a drug delivered to one patient affects the health of other patients in the population who are not treated.

consequence of a confounding structure that is very pervasive. As noted below, SES is just such a pervasive potential confounder in the case of many medical and social outcomes.

Hill seems to think of stability of association largely in terms of its evidential significance: it is one possible kind of evidence that an association can be interpreted causally. I argue below that there is something right about this, but that stability is most evidentially informative when an association holds not just in different circumstances but in circumstances that are sufficiently different that they are likely to involve different confounding structures. Of course, stability is not just of interest in its role of dealing with confounders-- causal relations that are stable are valuable for many additional reasons having to do with manipulation, control and explanation.

Stability and Mechanisms. What is the relationship between information about stability and information about mechanisms? The two are different although not unrelated. It is certainly possible to discover a stable causal relationship between *C* and *E* without knowing what mechanism connects *C* to *E*, as examples like smoking --> lung cancer and the aspirin--> headache relief illustrate. And knowing the mechanism connecting *C* to *E* may lead us to conclude that the *C*--> *E* relation is unstable/will not generalize well to a variety of circumstances if the mechanism is "special" and unlikely to be found in other circumstances. In addition, we may identify a mechanism connecting *C* to *E* in some circumstances and not know whether that mechanism is operative in other circumstances-- a mechanism found in mice may or may not be present and operative in the same way in humans. (Think of mouse models of neurodegenerative diseases.) On the other hand, when we have evidence that *C* causes *E* in one set of circumstances and are wondering whether that relation generalizes to other circumstances, information about the mechanism operative in the first set of circumstances can sometimes help to provide evidence (including evidence of a negative sort) about generalizability. To take an example discussed below, if the mechanism underlying causal relation between poverty and children's mental illnesses has to do with the increased time more affluent parents have for child supervision, increases in parental income that are the result of both parents working more than full time (as opposed to income increases from a stipend) would not be expected to benefit offspring mental health.

Stability in Relation to the Minimal Conditions for Causation. My discussion so far has adopted an analytical separation between whether a causal relationship holds at all in certain local circumstances (whether it satisfies the minimal condition for causation) and whether it generalizes to other circumstances (stability). Such a separation is assumed, either explicitly or implicitly, in many discussions of causal inference. For example, it is assumed in the contrast frequently drawn between internal and external validity: an RCT might establish that a causal relation holds between *X* and *Y* in a particular experiment (or perhaps in a population if the subjects are representative of the population) and thus that the experiment is internally valid in the sense of establishing causality in that particular context. However, it is taken to be a further question, not settled by the experiment, whether *X* will cause *Y* in other circumstances-- whether and to what extent the experimental result is "externally valid". A similar contrast holds for other causal inference techniques such as use of instrumental variables. Here current practice generally assumes that such inferences can establish (at best) only locally valid causal conclusions, holding for specific populations or even subpopulations, as in talk of "local average treatment effects" (LATEs) which are taken to hold only for those members of a population whose behavior is

changed by the treatment. (e.g., Imbens and Angrist, 1994) The extent to which such conclusions generalize is a distinct issue.

A related set of assumptions, perhaps not so frequently discussed, concerns trade-offs among different kinds of mistakes. The internal/external separation discussed above can be motivated by (something like) the assumption that it is better to avoid mistakenly claiming that a causal relationship exists when it does not than to fail to discover a causal relationship that does exist. In particular, if the kinds of causal inferences that are most reliable are those (such as RCTs) only establish local causal conclusions, it might be argued that it is preferable to employ these even if they don't yield results about the extent to which causal claims generalize to other contexts. The latter, it may be thought, are inevitably more risky and more likely to be erroneous.

This contrast between whether a causal relationship holds locally and the extent to which it generalizes is certainly defensible in some respects. With techniques like RCTs and instrumental variables, we have analytical results that establish that if the various conditions required for their correct implementation hold, claims about local causal conclusions (or more precisely claims about their expected error rates) follow deductively. Of course, we may mistakenly think that the required conditions hold when they do not and this may lead to invalid inferences but in such cases it can be argued that we have known sources of inductive risk. By contrast, generalization to other circumstances appears to raise other unknown and less controllable forms of inductive risk. Moreover, the contrast seems to track (to at least some extent) what might be described as one natural order of causal discovery, at least in portions of biomedicine and the behavioral and social sciences: First one establishes that a causal relationship holds locally, then (in a further step) one explores to what extent that relationship holds in other circumstances. If one does not have good reason to suppose that the relationship holds in some local circumstances, it seems pointless to worry about the extent to which it generalizes.

Nonetheless, the common assumption that the two steps (internal vs external validity etc.) are not just analytically separable but can be carried out completely independently of one another can be misleading and methodologically detrimental, as has become apparent in recent discussions of RCTs and instrumental variables. One basic problem is that a purely local result about causation, even if apparently internally valid, can be scientifically uninteresting and difficult to interpret unless accompanied by at least some information about stability or generalizability. Put in terms of our earlier discussion, too much focus on internal validity puts too much weight on the dependency aspect of causation (attempting to establish that *C* is a difference-maker for *E* in specific, local circumstances) at the neglect of the sufficiency or stability aspect. A closely related point is that sometimes evidence relevant to establishing a local causal claim can come from information about the apparent stability or generalizability of that claim to other circumstances, so that at the level of evidential reasoning, the internal and external aspects of validity are not always sharply separable. I will provide illustrations of these points below.

6. Causal Inference. Design- Based vs. Non-Design-Based

So far I have focused on distinctions among several causal concepts and some other features causal relations may possess. I turn now to a discussion of some strategies for causal inference. Given an interventionist account, a natural way of conceptualizing problems of causal inference

follows: First, in some (ideal) cases we may be able to carry out an intervention experiment of the very sort that is involved in our characterization of the relevant causal notion: For example, if we are interested in whether *C* is a net cause of *E*, we may intervene to change the value of *C*, observe a change in the value of *E* and conclude that *C* causes *E*, this inference being straightforwardly warranted because of the connection between intervention and causation. If we are interested in the average causal effect of a drug, we may be able to perform a randomized experiment. As we will see below, even in these cases we may be able to conclude much less than what we would like to know but the connection to the causal conclusion is pretty transparent.

Suppose, however, as is very often the case (both in psychiatry and elsewhere) we don't have data that is the result of a deliberate experiment-- instead we have "observational" data consisting of associations or correlations among various measured variables. Within an interventionist framework, a natural way of conceptualizing the inferential problem we face in these circumstances is that we are trying to figure out what the results of some relevant intervention experiment would be, were we to perform it, but without actually doing the experiment. This suggests (I don't say that it establishes) that we should look for data that is produced in a way that is "close to" or "emulates" data that would be produced by an intervention experiment (that is, data that is produced by an intervention-like process), the idea being that to the extent we can find such data, it is more likely to lead to reliable causal inferences than data that is not (known to be) produced in a way that emulates an intervention - like experiment.

There are a number of inferential strategies that have this general feature, to a greater or lesser degree. For example in the case of instrumental variables, the idea is to find a naturally occurring variable *X* that is related to some candidate cause *C* for *E* in the so-called soft-intervention-like way illustrated in figure 6: if (i) *X* is correlated with or known to affect *C* and (ii) in such a way that any variation in *E* that it causes occurs only though variation in *C* and not in some other way (this is known as the *exclusion restriction)* and an association between *C* and *E* is observed, we may conclude that *C* causes *E*. Importantly, this conclusion can be warranted even if there is an unobserved or unknown confounder *U* of *C* and *E*.

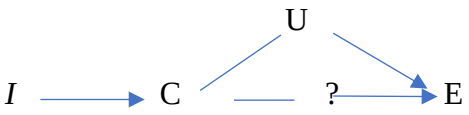


Figure 2 . The undirected edge involving *C* and *U* indicates that *C* and *U* are correlated and hence that the possible *C /E* causal relation is confounded (at least to some extent) by unobserved *U*. Adopting the Rubin-style counterfactual notation according to which E_c stands for the value *E* would have if *C=c*, *I* will be a valid instrument if (i) *I* _|_ E_c and (ii) I _/|_ C where _|_ stands for independence and _/|_ for dependence.

Other designs that can emulate interventions to a greater or lesser degree include regression discontinuity and differences -in - differences designs. In these designs too, the idea is to look for processes involved in the production of data that are intervention-like and to exploit this fact in causal inference. In these designs the result is virtually always a causal conclusion which is "local" in the sense that (just as is the case when one conducts an experiment) the upshot is a claim about the existence (or quantitative estimate) of a causal relation between a variable and an effect in some particular set of circumstances.

The strategies just described contrast with alternative strategies that are more global in the sense that they involve using the associations between a large number of observed variables to estimate causal relations among all of these at once, rather than focusing, as with the previous designs, on observations involving intervention-like designs that have as their target whether some particular causal relationship holds between two candidate variables. Typically, this second strategy involves trying to measure various potential confounders for causal relationships and then to eliminate their influence by statistical means-- e.g., by conditioning on them. I will accordingly call these conditioning strategies. These might be carried out by various statistical packages or machine learning algorithms. The result typically takes the form of a set of structural equations and/or a directed graph. Path diagrams such as the multi-level diagram representing causes of depression in Kendler and Prescott, 2006 are one example of this.

For both approaches--conditioning and intervention-emulation based -- control of confounders is crucial. However, the two approaches employ different strategies for accomplishing this. The intervention-emulating strategies are *design-based*. What I mean by this may be illustrated by a (rather idealized) physics example. Suppose one wishes to determine whether a certain particle X exists. When this particle is present, it produces a certain characteristic "signature" in a detector. Unfortunately, there are other particles *Y* that may also be present and can produce the same signature, making it unclear whether the signature in question is really due to the *X*s . So we have to control in some way for the presence of the *Y*s. Consider two possible strategies for doing this. One consists of trying to measure how many *Y*s are present or perhaps constructing a model that estimates how many are present and then using this to correct for possible confounding. A different strategy is to build a large shield around the apparatus that blocks all incoming *Y*s. Assuming that there are good reasons to think the shield is effective, there are obvious reasons for thinking that the second strategy is likely to be a more reliable. The first strategy is subject to the worry that we may fail to detect all of the *Y*s or we may have the wrong model of their behavior. The second strategy does not require that we are able to detect the *Y*s or have the correct model of their behavior-- the physical design of the experiment eliminates the possibility that they will be present. (Of course, we may be mistaken in thinking that the shield is effective but there will be independent ways of checking this-- e.g., turn off the experiment and see if in the presence of the shield there is any evidence for *Y*s.) In this sense the second strategy is design-based and the first isn't-- the first tries to achieve by measurement and calculation what the second achieves by the physical structure of the experiment.

A similar contrast applies to the two approaches to causal inference described above. In a randomized experiment, the physical design of the experiment itself-- the fact that there is random assignment -- can, when correctly implemented and additional assumptions are satisfied, ensure that the treatment and control group are equivalent in expectation with respect to other

factors that might influence the outcome[2]. Carrying out the randomization does not require that one know what these potentially confounding factors are or that one can measure them or correct for them by statistical means -- just as is the case with the physics experiment with the shield. A similar point holds for the instrumental variable strategy-- again, one does not have to know whether the potential confounder *U* in Figure 2 exists or be able to measure it, to be able to use this design to estimate the effect of *X* on *Y*.

By contrast in the more global, non-designed based approach we follow a strategy like that in the first physics experiment -- measuring the confounding factors and somehow correcting for their influence by calculation. There are various ways this can go wrong, the most obvious being that we fail to measure and correct for all of the confounders. But there are also more subtle problems with this strategy, including conditioning on the wrong variables (as opposed to failing to condition on the right ones[3].) In many cases, even on very optimistic background assumptions, such strategies will return at most an equivalence class of causal structures rather than a single unique structure.

These remarks are intended to supply some intuition for why one might think that the results of the design-based approach are often more reliable than the results of the non-designed based approach. But as I said above, I take them to be at best suggestive: as I see it, the choice between the two approaches is ultimately an empirical one-- which approach gives more reliable results in the sense of correctly identifying which causal relationships are present. Such empirical assessments can be accomplished in a variety of ways, including determining the ability of the different strategies to recover causal relations that are known on other grounds (a calibration strategy) , and the use of "triangulation" in which different strategies are assessed on the basis of whether they yield coherent or consistent results-- see below.

The recent literature seems to have largely settled on the conclusion that, as an empirical matter, the design-based approaches tend to produce more reliable results than the conditioning strategy. When the latter strategies are compared with the design-based, there is very often evidence that the former have not been successful in correcting for all confounders, different design-based strategies commonly yield more consistent results than the non- design-based strategies and so on. Randomized experiments often yield different results from purely observational studies, as do designs employing various instrumental variable studies. (Davey Smith & Hemani, 2014, Lalonde, 1986.) In econometrics and statistics, the increasing preference for design-based strategies is often described as the “credibility revolution", on the assumption that earlier, non-designed based causal inferences were in many cases simply not credible and design-based inferences are at least more credible. In what follows, I am going to accept this conclusion about the superior reliability of design-based approaches (in comparison with the alternative just based on conditioning) and for this reason will mainly focus on them. Having said this, however, let

[2] Of course as frequently pointed out this equivalence in expectation does not guarantee that in any particular randomized experiment this equivalence will be present.

[3] For example, if one is trying to determine whether X causes Y and conditions on a third variable Z which is an effect of X and Y (entirely possible if one does not know the correct causal structure), this will induce a conditional association between X and Y, making it look as though there is a direct causal relation between them

me also emphasize, as will become apparent below, that non-design based sources of information can also play an important role in securing reliable causal inference. Indeed, in many cases, the most convincing support for a causal claim will involve a process of triangulation in which mutually reinforcing evidence from a variety of different sources is brought to bear.

An Evidence Hierarchy? Readers will no doubt be familiar with the notion in evidence-based medicine (EBM) of an "evidence hierarchy" with (in some formulations) RCTs functioning as the "gold standard" at the top of this hierarchy. Like a number of other commentators, I don't find this picture entirely satisfactory, at least for the kinds of issues addressed in this essay. As suggested above, RCTs have a number of virtues, but they are certainly not the only source of information that legitimately can be used in causal inference. This is fortunate since there are many causal questions that cannot be addressed by RCTs, both for ethical and other (e.g. practical) reasons. The hierarchy picture also seems to inadequately stress the role of triangulation of evidence from different sources and theory elaboration discussed below. I do think, however, that there is something right about EBM's discussion of different sorts of evidence: different inferential strategies are subject to different sorts of errors and, relatedly, are good at providing certain kinds of causal information and not others. It is a good idea to be cognizant of these differences, even if they don't translate into a hierarchy.

7. Randomized Controlled Experiments and their Limitations.

As suggested above, although design-based inferences are generally more reliable than a pure conditionalizing strategy, they nonetheless have a number of important limitations. I turn now to a discussion of these, beginning with RCTs.

As I understand RCTs, they are designed for a particular kind of problem situation, although one that is common in parts of the biomedical and behavioral sciences. The situation is this: one has a candidate cause *C* and effect *E* but it is also known that the units or systems to which *C* may be applied differ among themselves in all sorts of other unmeasured ways that may be causally relevant to *E*. In particular, the units may be causally heterogeneous and features of the units may be differentially correlated with causal factors besides *C* that affect *E*. Randomization attempts to address this problem by creating two groups that do not differ in expectation with respect to *E*, except for the fact that one group receives *C* and the other does not. As remarked above, there are many experiments in which this is not necessary-- one may be confident one is dealing with homogenous units or one may be able to remove the in-addition - to - *C* influences on *E* by other means, shielding, preparing pure samples and so on. Randomization is what one does when these other strategies are not available. It does deal with unknown confounders, at least in expectation, but its employment comes with costs.

Of course it is also true that an RCT can be badly designed -- e.g., there may be failure of double blind conditions as when subjects can tell that they have received the active drug and not a placebo, employment of a defective outcome measure, fail to account for non-compliance, violations of SUTVA and so on but let's put these possibilities aside and consider what can be learned from a well-designed RCT that does not have these sorts of defects. Even in this case some important limitations remain, many of which have to do with the fact that RCTs don't provide evidence about the stability of causal claims, in several different senses of stability.

What is the Active Ingredient? Consider the following hypothetical randomized experiment drawn from Esterling et al.(n.d.) Researchers induce the Riverside (CA) Superior Court to mail postcards with an official government seal to residents who have received a jury summons, randomizing so that half receive a standard reminder postcard and the other half a postcard informing recipients that failure to appear can result in fines and imprisonment. This second "enforcement condition" results in a statistically significant increase in turn-out in comparison with those in the control condition. The researchers conclude that enforcement messages cause increases in jury turnout. Next, they repeat this experiment in Orange County (a nearby but more affluent location). To their disappointment, in this sample, there is no effect of the enforcement message. In an effort to figure out what is going on, the researchers resolve to repeat their original experiment in Riverside County. This time, however, they do not have the cooperation of the Riverside court so they send out randomized postcards with the same enforcement message but without the official court seal. Again, there is no treatment effect.

Let's stipulate (what may not be warranted) that the original RCT produced a valid result in the sense that it captures the ACE of the original manipulation performed on the Riverside population. One way of putting the problem posed by this example is that while the researchers can conclude that, in the first experiment, they performed a manipulation that had the observed effect, their result does not tell them what was the "causally active ingredient" in that manipulation. That is, they don't know what features of their manipulation were causally relevant to the outcome they observed and which were irrelevant in the sense that these could have been omitted and the same result would have ensued. Another way of conceptualizing the problem is that what is actually known to be manipulated in the Riverside experiment (whether or not subjects receive a certain message with the official seal of the Riverside superior court about penalties for failure to appear for a jury summons) is not the same as what is assumed to be manipulated when the results of the experiment are described in terms of the generalization "enforcement messages increase jury turnout". The former (the message from the Riverside Superior Court) is (at best) a particular way of operationalizing the more general category of "enforcement message" which the generalization takes to be the relevant causally active ingredient. Given the Orange County results and the second Riverside experiment, the original Riverside result does not support this more general claim about the effects of "enforcement messages". The latter is a claim about what is stable in the Riverside result (that it is presence of an enforcement message per se that stably increases turnout). Even though the original Riverside experiment apparently found an ACE, this experiment does not tell us anything about the stability of that result and this failure is sufficiently acute that it makes the experiment difficult to interpret.

In some cases in which an RCT is performed, one has background knowledge or good grounds for belief, coming from theory or some other source about the active element in the manipulation, if there is one. If the treatment involves ingestion of a pill, and there is adequate control for placebo effects, one may have good grounds for assuming that it is the material in the pill that is causally relevant to any observed outcome, rather than, say, variations in the manner of swallowing. (Of course, it is possible to be wrong about this and even here the RCT by itself will not establish exactly what in the pill was relevant to the outcome.) The (or one) problem we face in the jury summons example is that we don't have anything analogous to even the weak background knowledge we have in the pill example. As a consequence, the RCT is difficult to

interpret and of limited usefulness. As this example illustrates, although part of the appeal of an RCT is that it requires very limited theory or background knowledge, it does require some to be useful.

Ambiguous Interventions/Heterogeneous Effects. I turn next to a second illustration of a limitation of RCTs, related to but not the same as the first. This involves a much discussed example in the causal modeling literature due to Spirtes and Scheines, 2004. Following them, I'm going to distort the history and relevant science a bit to make the illustration more salient. Assume that total cholesterol (TC) is simply the sum of high density lipoprotein HDL and low density lipoprotein LDL. Initially these two forms of cholesterol were not distinguished and were measured by TC which was taken to causally contribute to heart disease (HD). It was subsequently discovered that HDL was protective for heart disease while LDL had an opposite, deleterious impact.

Now consider a randomized experiment in which TC is set to different levels in a treatment and control group. The association between TC and HD will depend on the precise mix of HDL and LDL that realizes TC in this particular experiment. If the imposed TC is mainly made up of HDL, the impact of TC on heart health will seem to be positive for greater levels of TC. If the TC is mainly made up of LDL the impact will seem to be negative. If the mixture of HDL and LDL realizing TC across different experiments varies, the association between TC and HD across different experiments (and hence its apparent causal impact) will be variable or unstable and, arguably, not well-defined.

Intuitively, the problem with TC is that is not a causally homogenous variable-- instead it is a mixture of two different variables with very different causal effects. At an abstract level this is another example of not knowing what one is manipulating. However, it has a different structure from the Riverside example. Conducting a single RCT with TC will not tell us that this heterogeneity is present, although if a number of RCTs are conducted with different mixtures of TC, the resulting instability of the results may be suggestive of heterogeneity.

In the jury summons example, we were not sure what the active ingredient was, but we assumed, perhaps wrongly, that anything that was manipulated that did not contribute to the observed result was neutral or irrelevant to it. In the TC example, the problem is that there are two ingredients in the overtly manipulated variable that have "opposite" effects. In both cases, however, the problem is that we lack the right sort of information about what it is that we are manipulating. As these examples illustrate, causal inference, even when design-based, can lead us astray when we have the "wrong" variables.

To what extent do candidates for causal variables in psychiatry exhibit problems of the sort described above? I'm not competent to provide a general answer to this question but one obvious case in which similar problems arise has to do with RCTs that attempt to evaluate various forms of psychotherapy. There are a well-known number of methodological problems that arise in the use of RCTs for this purpose, but one issue is this: the RCT needs to specify a protocol for treatment that (one hopes) will be followed uniformly by all therapists in the treatment arm of the experiment. In such cases even if there is an apparent positive average causal effect, the RCT may leave it unclear which features of the protocol causally contributed

to the beneficial result and which were causally irrelevant, in analogy with the problem with the social science experiment discussed above[4]. This of course creates problems for generalizability to other contexts, particularly since it is unlikely (very difficult to ensure) that the protocol will be exactly repeated in those contexts. Again, this might in principle be addressed by additional RCTs that vary features of the protocol but this may be difficult and expensive. One might hope that over some range of variations in the protocol the beneficial effect will be roughly the same (or will at least have the same sign in the sense of being at least somewhat beneficial) but whether this is so is of course an empirical question[5]. This issue is discussed by Shrout (2011) who considers the use of mediation analysis to identify the active ingredients in cognitive behavioral therapy (CBT), one of the most popular and well-studied forms of psychotherapy. I will say a bit more about this below.

Next consider that many of the environmental variables thought to influence mental illness (whether these are analyzed through design- based inferences or in some other way) may be subject to varying degrees to the problems described above. Low SES is associated with increased risk of many mental illnesses, and it seems plausible that there are specific ingredients in low SES that are causal for these outcomes. On the other hand, depending on how it is measured, SES will be a mixture of many different more specific variables having to do with income, education, cultural capital and much else. Some of these may be more causally relevant to mental health than others (and different mixes of these may be present in different subjects with same SES) and this may affect the generalizability of results. Moreover, the various components vary in the extent to which they can in fact be manipulated or even whether such manipulation makes clear sense. Income can be manipulated-- hence it can be a treatment in an RCT and might be identified as a cause of some mental disorders in an instrumental variable analysis, as in the Smokey Mountain study described below, but it is less clear what might be involved in an intervention on cultural capital or, for that matter, education[6]. Moreover, a skeptic might wonder whether even income is a causally homogeneous variable on the grounds that income derived from a grant, or a lottery may have different causal effects than "earned" income-- again see the Smokey Mountain study. In general, when the value of a variable is somewhat under individual control or reflects individual choices, one might wonder whether it has the same causal properties as a value of that variable that is exogenously imposed. (In addition to receiving a stipend vs earned income, contrast chosen vs imposed school attendance,

[4] A more alarming possibility is that some elements of the protocol are deleterious for mental health, but their effects are often masked by advantageous elements, in analogy with the total cholesterol example. Perhaps, though, it can be assumed that usually the inactive elements are causally irrelevant.

[5] One might think of this as a kind of "monotonicity" assumption: if individuals respond to the treatment at all, they will respond in a "positive" way or their probability of a positive response will be boosted rather than diminished. This may be a reasonable assumption in many cases but course its warrant is ultimately empirical.

[6] "Assigning" subjects different educational levels is morally objectionable and in any case is very different from observing subject's educational levels, when these reflect subject choices, the resources they have available and so on.

or mandated psychotherapy vs voluntarily chosen psychotherapy.) To the extent this is so, it will be a complication in the interpretation of intervention-like designs.

Population Relativity. So far, I have focused on the fact that experimenters may not know what it is about their manipulation that is relevant to the outcome they observe. (The same is true for other designed-based strategies, as noted below.) However, there is another problem that at least in practice is difficult to separate from the first. This is that, as noted above, the notion of an ACE (which again is the upshot of an RCT and many other design-based inferences) is a population-relative notion and even if the subjects in such an experiment are randomly sampled from some clearly defined population, strictly speaking the RCT will only provide information about that population[7]. For example, it might be that the reason why the original jury summons experiment (performed in Riverside) failed to replicate in Orange county doesn't have to do with the variation in enforcement message employed but rather occurs because the residents of the latter are, on average, causally different in some way from the former-- e.g., the latter are more affluent than the former and more affluent people are less affected by enforcement messages. The problem is that we can't tell from the failure of replication in Orange County whether this results from a causal difference in the two populations, or the variations in the manipulation performed in the two experiments or both.

The Significance of Individual Variability. In the total cholesterol example, TC had different "effects" (arguably not well-defined effects at all), depending on the mix of LDL and HDL. We did not, however, consider the possibility that different subjects might have different responses to LDL and HDL-- instead we assumed that the level of these had the same constant effect on each subject. However, another possibility is that the treatment variable in an RCT has different effects on different subjects in a way that cannot be captured in a constant effect representation[8]. The result of an RCT is an estimate of an average effect, and this average is consistent with substantial variability in individual response. Of course, such variability is very often observed in both the treatment and control group and the RCT by itself provides no insight into why this variability occurs. A particularly disturbing possibility in biomedical contexts is that although the average effect of the treatment across the entire treatment group may be positive, some subjects may be harmed by the treatment while a larger group benefits. Rothwell (1995) describes a study of the results of carotid endarterectomy (a surgical procedure that removes plaque from the

---

[7] The issue of whether there is random assignment in an experiment is of course distinct from whether the participants are a random sample from some population. In some (many?) cases the participants in an RCT represent a sample of convenience that is not randomly drawn from any well-defined population. For example, drug company run RCTs are typically unrepresentative of patient populations. This raises additional problems of generalizability.

[8] A standard way of modeling individual variation in an RCT represents this as due to the distribution of an additive error term U around a constant effect : $Y = f(T) + U$, where T is the treatment with constant effect given by $f(T)$ and individual variability is due to different individuals having different values of U. Note that this does not seem to capture what is going on in the stroke treatment example immediately below, where there is nothing corresponding to this constant effect. In many if not most cases in which there is variability of individual response to treatment according to differences in individual "type" this is not captured by an additive error term.

carotid artery). Although this procedure reduces the risk of stroke for many patients, it can also increase the risk of stroke by dislodging plaque that can contribute to arterial blockage. Rothwell employed a "prospective" model that accurately predicted stroke risk in the patients studied. He found from this that surgery reduced the risk of stroke in those patients at high risk of stroke, but appeared to increase the risk of stroke among low- risk patients.

Of course, there is also a great deal of variability in the responses of patients with mental illness to various drugs and therapies. The sources of this variability are in many cases not well understood and as noted above, are not addressed by standard RCTs or other designed-based strategies, which also just provide information about average causal effects. In the study described above, Rothwell had an empirically supported model which allow him to distinguish among patients according to the degree of stroke risk they faced. My impression is that in psychiatry finding markers that allow for the prospective classification of patients into subtypes depending on how they are likely to respond to a treatment has turned out to be difficult, although I assume genotyping or the use of PRSs may afford some possibilities.[9] Conducting retrospective analyses in which one looks for distinguishing characteristics among patients who have responded differently to a treatment carries obvious risks of overfitting. Understanding variability is of course beneficial for individual patients but in addition it bears on issues of stability and generalizability-- if different types of subjects respond differently to treatment and these types are distributed differently across different populations and circumstances then even average effects will be unstable.

Note also that the problems described so far (lack of information about the active ingredient, population relativity, subject heterogeneity) are not problems of "confounding" in the ordinary sense. (Again, I take confounding to arise when there is a non-causal association between *X* and *Y* which is due to, e.g., some omitted variable and which we mistakenly interpret as causal). Instead, the problems are best conceptualized as problems arising from lack of stability/invariance in the identified causal relationship or an inappropriate choice of variables. These problems illustrate the point that RCTs and other design-based inferences can address problems having to do with "local" confounding, at least in principle, but they do not address the other problems discussed above.

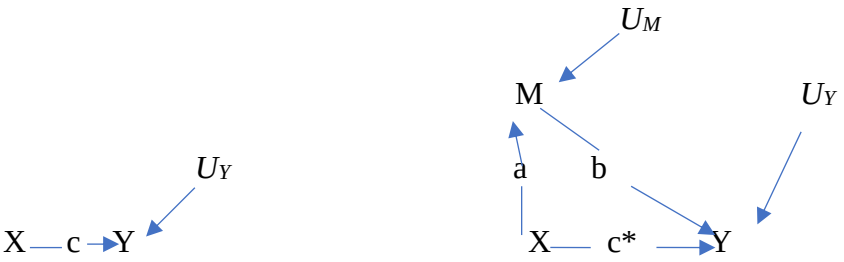
8. Total vs Path-Specific Effects Again

I turn now to another limitation of an ordinary RCT (and other intervention-emulation experiments). Even if otherwise unimpeachable by themselves, designs of this nature tell us at most about the total causal effect of the candidate cause and do not provide information about distinct paths by which this cause may influence the effect. Recall our earlier example of a gene *G* that affects an outcome *T* via an inside-the-skin pathway involving neural processing but also via a mediating outside-the-skin pathway that goes through the external environment and the subject's behavior *B*, as when *G* also causes a tendency to explore intellectually rich environments or, in the case of depression, a tendency to behave in a way that elicits negative

[9] There have been a number of studies that attempt to do this but Kenneth Kendler notes (personal communication) that "progress in pharmacogenetics has been slower than expected - better at predicting side effects such as weight gain, than therapeutic effects".

events which further contribute to depression. In such cases a single experimental manipulation of *G* will reflect its total effect on *T* summed over both these paths. Of course, if we know where to look for the outside-the-skin variable, we could in principle perform a second intervention which fixes the value of that variable while at the same time manipulating *G*. This would separate the two paths and reveal that portion of the impact of *G* on *T* which is direct or inside the skin.

Mediation Analysis. In psychiatry and psychology, the much more common case is (at best) one in which there is randomization of/intervention on the treatment variable but no intervention on or independent randomization of any candidate mediating variable (that is a variable that is on a path in between the treatment and its putative effect) that might be used to distinguish different paths. Instead, the value of any candidate mediating variable (and its association with the treatment and candidate effect) is merely observed. In this situation, *mediation analysis* is sometimes performed to attempt to understand the causal role, if any, of the mediator. Without going into a lot of statistical detail, consider an early version of this idea which was developed by Baron and Kenny (1986). Suppose *X* has a net or total effect on *Y* and that this may be at least partially mediated through an intermediate variable *M* where both *Y* and *M* have additional sources of variation-- U_M and U_Y -- and all of the relationships are linear. Suppose (very optimistically) that U_M and U_Y are uncorrelated. Suppose also (1) *X* is related to *Y* via coefficient *c*, (2) *M* can be measured and is found to be associated with *X* via coefficient *a*, and (3) after controlling for *X*, *M* is related to *Y* by coefficient *b*. (We estimate these coefficients by ordinary regression analysis.) We then proceed by estimating the "direct" effect *c** of *X* on *Y*, controlling for *M*. If $c^* < c$ this shows the relationship between *X* and *Y* is at least partially mediated by *M*.



Although this design can, under the assumptions above, lead to the identification of a mediating variable (a mechanism or part of a mechanism?) linking *X* to *Y*, the assumptions are highly restrictive and are likely often violated. In particular (and even putting aside the assumption of linearity) since *M* is not randomized or the result of an intervention, the design does not exclude the possibility that relation between *M* and *Y* is confounded (due to a correlation between U_M and U_Y) resulting in a biased estimate for the indirect effect. More recent work has shown how to relax the assumptions about linearity, but it remains the case that an accurate decomposition into path effects requires strong assumptions that are often not satisfied. A recent meta-analysis (Stuart et. 2021) showed that the majority of papers included did not satisfy the assumptions required for reliable mediation analysis.

Putting aside these statistical issues, mediation analysis can in principle be helpful not just with decomposition of total effects into path specific effects but with the problem of identifying active ingredients described earlier. Shrout (2011) draws attention to a randomized controlled experiment conducted by Freedland et al (2009) which found that CBT is associated with reduced symptoms of depression *D* among patients who have experienced coronary artery bypass surgery-- reduced in comparison with an alternative treatment for depression or no treatment at all. As Shrout observes, one might wonder what "component" ("active ingredient") of CBT was responsible for this effect-- (1) control of (*T*) distressing automatic thoughts or (2) control of (*A*) dysfunctional attitudes. (Assume that there are theoretical considerations that support the claim that *T* and *A* are distinct and that these are the two most likely mediators.) If we could measure *T* and *A* and their association with whether subjects received CBT and the incidence of depression and the relevant requirements were satisfied, we could use mediation analysis to determine whether *T* or *A* was a mediating variable. This in turn might help to improve the efficacy of CBT and its generalizability as a treatment for depression in other circumstances.

9. Instrumental Variables.

Kendler et al. 2018 provide an interesting and in my view convincing example of the use of instrumental variable analysis in a psychiatric context. Poor academic achievement *AA* has long been known to be associated with drug abuse *DA*. However, this association could arise from several potential confounders-- for example, family background might act as a common cause of both *AA* and *DA*. Kendler et al. used a complete population cohort of Swedish students aged 15 to 20 to investigate a possible causal relationship between *AA* and *DA*. Sweden, like many other countries, has a cut-off date for school enrollment with the consequence that students who are born in the same year are placed in the same class-- thus students in the same class can differ in age by as much as 12 months. Unsurprisingly, the younger students in a given class tend to do less well in terms of *AA*. Kendler et al. used student month of birth (MOB) as an instrument. Their assumption was that MOB affects *AA* (again there was considerable independent evidence for this) and that it affects *DA* if at all only through *AA*. Without going into all of the statistical details, their IV analysis found a significant association between *AA* and *DA* underage-associated variation in *AA*, thus supporting the claim that *AA* has a causal affect on *DA*. They also found, reassuringly, that conditioning on *AA*, there was no association between month of birth and *DA*, thus supporting the claim that any influence of month of birth on *DA* occurred through *AA*. Although the authors do not make this observation, it is also relevant that if the direction of causation was instead from *DA* to *AA*, conditioning on *AA* should induce a conditional dependence between *AA* and month of birth and this was not observed.

The authors also employed a co-relative design, exploring the association between low *AA* and *DA* in first cousins, full-sibling and monozygotic twins who were discordant for *AA*. This provided varying degrees of control for genetic factors and family environment. They again found results that were consistent with their model of the influence of *AA* on *DA*-- that is, among monozygotic twins who are discordant for *AA*, the twin with lower *AA* was more at risk for *DA* and a similar pattern was found among other relatives. *AA* is a variable that might be intervened on and the authors also cite a meta-analysis showing that programs in the U.S and elsewhere aimed at improving *AA* in targeted students were also associated with decreased drug use, consistently with the authors' causal hypothesis. It is plausible that these alternative designs--

instrumental variables and co-relatives as well as direct interventions on *AA*-- have different possible sources of confounding. If the association between *AA* and *DA* is non-causal, different kinds of confounding would have to be present in each of the three designs that are "coordinated" in such a way that they happen to produce this consistent association between *AA* and *DA*. There is no reason to expect this. The convergence of the different designs on the same conclusion thus strengthens support for this conclusion. I will say more below about the benefits of this "triangulation" strategy.

As I said, my assessment is that these results provide good evidence that *AA* causally influences *DA* at least in some circumstances. But if we focus just on the instrumental variable part of the Kendler et al. study, there are obvious worries one might raise about the stability and generalizability of the result, again illustrating our general theme about the local character of the results of design-based inferences. In particular, as far as the instrumental variable part of the study goes, even if one accepts that the results support a causal claim about the population of Swedish students studied (i.e., that this causal claim is internally valid) one might be inclined to suspend judgment about the extent to which the claimed causal connection holds in other populations or circumstances. (Of course, the meta-analysis involving intervention experiments helps to address this worry.) More radically, one might conclude that we know only that the results hold for the particular cohort studied. Even more radically, some might think that we are entitled to conclude only that the *AA*--> *DA* result holds for students whose *AA* is affected by their age, rather than for all students. (If this seems unduly skeptical, similar claims have been made by prominent econometricians in interpreting the results of instrumental variable analyses.)[10]

In addition, although the study is framed in terms of the causal influence of *AA*, one might wonder about what the causally active ingredients of *AA* are. The authors measured *AA* by grade point average, but they noted that

> our measure of AA does not perfectly reflect the way this construct is typically defined in the education literature, which includes students' attitudes about their teachers, their commitment to school, their educational aspirations, and their level of truancy.

Strictly speaking, the study might thus be taken to show only that grade point average influences *DA* since that is the measured variable. On the other hand, there are good reasons for wanting to know whether the other elements of *AA* described above influence *DA*. For example, in some cases it may be easier to intervene to affect student attitudes and truancy than their grade point averages.

10. An Additional Role for Stability.

I noted above that sufficiently radical failures of stability in a putative "local" causal relation can undermine its interpretability. But stability considerations can be important in another, more

[10] One might also wonder whether month of birth satisfies the exclusion restriction. Suppose that younger students are more socially immature *S* and that *S* influences *DA*, either positively or negatively, via a path that does not go through *AA*.

positive way: as a device for detecting when confounders are present and providing some support for claims that they are absent. The underlying intuition is that if a putative causal relation seems to continue to hold across (what are plausibly regarded as) different potential confounding structures, this supports the claim that the relationship is genuinely causal. By contrast if the relation fails to hold across different potential confounding structures or if it changes considerably in strength, this may undermine the claim that it is causal. In employing this form of argument, it is important that the potential confounding structures are genuinely different. As observed earlier, it is a very real possibility that an association might continue to hold across what seem to be "different" circumstances but nonetheless be non-causal-- the association instead holding because it is generated by an underlying confounding structure that is present across all those circumstances. In biomedicine and the behavioral sciences, confounding structures, due to pervasive and entrenched variables like SES, can have this feature.

An illustration of this strategy is provided by an investigation into the effects of maternal smoking on children's outcomes. In previous studies, maternal smoking was found to be robustly associated with conduct and cognitive problems, and hyperactivity as well as low birth weight (lbw). Sellers et al. (2020) compared the association of maternal smoking with outcomes in two different UK cohorts-- one born in 1958 and the other in 2000-1. The authors found the association with lbw was relatively stable across these two cohorts. By contrast, the association with conduct disorder etc. was considerably stronger in 2000-1 than in the previous cohort. At the same time smoking in general decreased from 1958 to 2000-1 and became much more strongly associated with low SES and social disadvantage. This suggests, even if does not conclusively establish, that the association between maternal smoking and conduct disorder etc. is confounded by SES effects[11]. The contrasting stability of the relation between maternal smoking and lbw across these demographic changes makes it more plausible that this is a genuine effect-- perhaps especially because the effect, if present, is plausibly understood to be the result of biological processes that operate independently of social circumstances[12]. Results from other studies support this conclusion via triangulation arguments. For example, behavioral and cognitive problems in offspring but not lbw are associated with paternal smoking, again suggesting that social disadvantage is casually contributing to these problems, with paternal smoking being correlated with social disadvantage. The absence of a correlation between paternal smoking and lbw is what one would expect if maternal smoking was genuinely causal for lbw. In other studies involving monozygotic female twins who are discordant for smoking, lbw of offspring is observed in the smoking twin in comparison with the non-smoker, suggesting that this effect that is not entirely due to genetic or shared family background. Other possible analyses might compare the birth weights of siblings whose mother smoked during one pregnancy but not the other -- again if the effect is causal, one would expect a difference in birth

[11] Kenneth Kendler has remined me that this issue has also been studied using co-relative analyses and mothers who smoked in one pregnancy and not the other-- another example of triangulation. Both approaches suggest that the association between smoking during pregnancy and adhd in offspring is not causal - see, e.g., Skoglund, C. et al. (2014).

[12] This involves an assumption about a likely mechanism but note its *conditional* form: if a certain generic sort of mechanism is present, we would expect a certain pattern of results. This is different from establishing that such a mechanism exists and then using it to support a causal conclusion. See below for additional discussion.

weights across the two cases. Yet another technique appeals to Mendelian randomization, using as an instrument a genetic variant associated with greater cigarette consumption among those who smoke. Mothers who have the variant and smoke had offspring with lower birth weights than non-smokers with the variant. Here the variant acts as an independent, random source of variation in smoking behavior--- finding an association between this variation in smoking and lbw strengthens the claim that this association is causal.

11. A Role for Causal Specificity.

Presumably no one doubts that it is desirable to discover stable causal relationships when they exist. The notion of causal specificity has a different status -- many have thought that whether a cause is specific has no special significance for reliable causal inference. As I noted above, Bradford Hill did not agree. He lists specificity as one of the "aspects" of causation that is relevant to establishing a causal conclusion. He illustrates the notion in the context of diseases associated with working in certain industries, as follows[13]:

> If, as here, the association is limited to specific workers and to particular sites and types of disease and there is no association between the work and other modes of dying, then clearly that is a strong argument in favour of causation.

This suggests the following general characterization: a causal relation involving a cause *C* is specific to the extent that *C* is associated with a single kind of effect or a small number of kinds, rather than many different kinds of effect[14]. That is, *C* is non-specific to the extent that it not only causes (or appears to cause) E_1 but also to cause the different kinds of effects E_2, E_3 and so on. In Hill's example, an exposure associated with working in a certain industry will be specific if it is associated only with a particular kind of cancer (or a small number of such kinds) and less specific if is associated with death from many other diseases.

Although Hill clearly thought that the specificity of an association was relevant to whether it was causal, he noted that some genuine causal relationships can be relatively non-specific-- for example, smoking causes many different diseases. More recently a number of epidemiologists

---

[13] Hill's running example is cancers among nickel refiners. Immediately after introducing the characterization of specificity quoted above he writes: "We must not, however, over-emphasise the importance of the characteristic. Even in my present example there is a cause-and-effect relationship with two different sites of cancer – the lung and the nose. Milk as a carrier of infection and, in that sense, the cause of disease can produce such a disparate galaxy as scarlet fever, diphtheria, tuberculosis, undulant fever, sore throat, dysentery and typhoid fever". Similarly cigarette smoking causes many different kinds of cancers but this does not undermine its status as a cause of each of these. Thus, as Hill goes on to remark, while the presence of specificity suggests (although it does not conclusively establish) the presence of a causal relationship, its absence need not undermine the claim that a relationship is causal. As I observe below, one way of thinking about this is that when specificity is present, this can suggest the absence of confounding.

[14] What is it for effects to represent "different kinds" rather than a single kind? This seems to involve context specific judgments of various sorts that I will not try to elucidate.

have been completely dismissive about the role of specificity in causal inference. For example, Rothman and Greenland (2005) describe specificity as a "wholly invalid" consideration.

Against this, I follow several recent discussions (e.g., Blanchard, 2022) in holding that specificity-like considerations can be relevant to the reliability of causal inference. Basically, this is because specificity in a causal relationship can sometimes support claims about the absence of confounding and lack of specificity can suggest that confounding may be present. Weiss (2002) gives the following example. Suppose a lower rate of head injury is observed in cyclists who wear helmets as opposed to those who do not. One interpretation is that the helmets protect against head injury. However, there are possible confounds: perhaps those who wear helmets are more careful in general and thus less prone to head injury. In this case if the protection is specific to head injury-- that is, if there is no association between wearing a helmet and injury to other parts of the body-- this counts against the operation of a confounder of the sort described. On the other hand, if helmet wearers are also less likely to suffer non-head injuries this suggests that it may be general carefulness rather than their helmets that is responsible for their lower rate of head injury. Note that as far as this second possibility goes, it is not (on the reconstruction offered above) the mere fact of non-specificity that is relevant but rather that the non-specificity is of such a character as to suggest confounding. To compare this example with the case of smoking, one might think it plausible that smoking causes several different kinds of cancers (of the lung, esophagus etc.) perhaps because all of these organs are exposed to smoke, as well as heart disease, so finding this sort of non-specificity would not necessarily suggest confounding, in agreement with Hill's assessment.

It is worth noting that the helmet example might also be understood in terms of a mechanism-based reasoning: if helmets protect against head injury, the mechanism by which they do so involves covering portions of the head but not other parts of the body and this has implications of the sort described above. However, it is not clear that mechanistic information figures in the above inference in quite the way that proponents of a role for mechanism in causal inference sometimes suggest. These proponents suggest that to establish a causal relationship one must show both that an association is present and, independently, identify a mechanism that accounts for the association or, more weakly, at least provide evidence that such a mechanism exists. However, in the example just described, it is not known whether helmets are a mechanism that protects against head injury or even that a head-injury-protecting mechanism exists that explains the observed association -- instead this is what one is trying to find out. The logic of the reasoning is rather conditional: *if* helmets are a mechanism that protects against head injury, one expects one pattern of association; if they are not, one expects another pattern of association. Put differently, the role of the possible mechanism is to suggest additional predictions if the original causal claim is correct-- what is sometimes called "theory elaboration", as I note below.

Another illustration of the use of specificity in assessing whether confounders may be present is provided by the notion of "negative controls". Smoking has been robustly associated with suicide in a number of observational studies. This might be thought to support a causal connection. However, as shown by Davey Smith, et al.,1992 smoking is equally strongly associated with homicide-- that is, smokers are more likely to be homicide victims than non-smokers. Even if we suppose that there is some mechanism by which smoking causes suicide, it is very hard to see why this same mechanism should also lead from smoking to homicide

victimhood. Instead, it is far more likely that the smoking/ homicide association is due to demographic/environmental factors (people with lower SESs are more likely to smoke and also more likely to live in more dangerous neighborhoods) or other uncontrolled confounders. Since we have strong evidence that the smoking/homicide association does not control for such factors, it is arguable that this should undermine our confidence that such factors have been adequately controlled for in connection with the smoking/suicide association. Note that here too the reasoning is conditional insofar as it involves an inference about "mechanisms".

I remarked above that the mere fact that an association between observed variables is found in a number of different studies or background circumstances is not a very strong reason for taking that relation to be causal. The association may instead be due to a common cause or confounding structure which is present in all of these circumstances-- a possibility that is by no means unlikely when, for example, the confounding structure involves entrenched socio-economic or cultural factors. The smoking/ homicide example illustrates this[15].

12. Elaboration of Cause/Effect Relations/ Dose Response Relations.

As I see it, the above role for specificity is an instance of a more general strategy which can be employed in causal inference: that of *elaborating* a putative cause/ effect relation. Suppose there is an association between *C* and *E* and one wonders whether it is causal. Often one has additional information or grounds for belief about what that relation would be like *if* it is causal. (This information may be vague and qualitative, as illustrated below.) One can then sometimes use this additional information to help assess whether confounders are present that render the association non-causal.

A simple illustration is provided by the existence of a dose/response relation-- another of Hill's "aspects" of causation. Suppose one observes an association between smoking and lung cancer. A commonsense thought is that if this relation is causal, it must involve the passage of some material from the smoke into the lungs; hence it would not be surprising if the incidence of lung cancer increases for those who smoke more heavily or for a longer period of time or who report inhaling more deeply. In fact, these relationships are observed. But a conclusion that smoking causes lung cancer does not rely just on the fact that this hypothesis predicts what is observed. The observed dose/response relations have another role as well: they can help to make it implausible that the smoking/lung cancer association is entirely due to confounding. The reason for this is that for some confounding hypothesis to account for the observed associations

---

[15] Another role for specificity is this: We noted above that when an instrumental variable is employed in causal inference, it is crucial that it satisfy an exclusion restriction: the instrument should affect the putative effect *E* only through the putative cause *C* and not via some other route that does not go through *C*. To the extent that an instrumental variable (or for that matter, an ordinary intervention variable) is relatively specific, this provides some support for the claim that the exclusion restriction is likely to be satisfied. By contrast, if a purported instrument is highly non-specific, with many different effects, this raises a concern that at least one of these involves a path that does not go through *C* and yet affects *E*.

including the dose/response relationships, the confounders would need to take a very specific form or be distributed in a very precise way which we have no reason to expect. For example, if there is a common cause *X* of smoking and lung cancer (e.g., some genetic factor as suggested by Fisher) then *X* would need to be such that variations in it cause variations in how much people smoke and for how long and how deeply they inhale that are precisely matched to variations in their probability of lung cancer. This is not impossible, but it arguably makes the hypothesis of complete confounding less plausible.

Note that in arguing this way we are *not* assuming that for a relationship to be causal there *must* be a dose/response relation. (Hill is sometimes mistakenly interpreted as claiming this.) Rather the argument is that (i) it can be plausible to expect such a relation in some specific cases (which is consistent with its not being present in others), (ii) one can observe whether such a relation is present, and (iii) if it is present, one can appeal to this to help render implausible certain hypotheses about confounding. A dose/ response relation if present is just one way of elaborating or developing a hypothesis about a causal relationship in a way that may help to exclude alternative hypotheses about confounding.

I am not aware of appeals to a dose/response relation being widely used in causal inference in psychiatry but there do seem to be some examples and this is a strategy that might be more widely used. Costello et al. (e.g., 2003) report the results of a study (the Great Smokey Mountains Study) of a population of rural children, a quarter of whom were American Indian and the remainder white. The opening of a casino gave each Indian an annual income supplement. Consistently with findings from other populations, children whose families received the supplements showed reduced risk for problems such as conduct disorders. Moreover, younger children who were exposed to this increase in family income over a longer period of time showed a reduced risk in comparison with older children who benefitted from the supplement for a shorter period of time. In other words, a stronger exposure to the putative cause was associated with a stronger response in the form of reduced conduct problems. As argued above, this makes it at least somewhat less plausible that the observed association is due to confounding-- any confounders would have to be distributed in such a way that they are associated with both additional income and reduced conduct disorders in a way that reproduces the dose/response relation between the two.

The role of poverty and family income in children's mental health problems also provides additional illustrations of many of themes in this essay. Costello et al. (2003) report the results of tests of a number of possible mediators of the connection between poverty and mental health, including traumatic life events (e.g., parent separation or divorce, sexual or other physical abuse, unplanned pregnancy), neglect, harsh or inconsistent parenting, overprotective or intrusive parenting, lax supervision, and maternal depression. Of these only one met the requirements described above by Baron and Kenny for full mediation: failure of parents to provide adequate supervision, which accounted for 77% of the effect of changing poverty level on symptoms. As noted above the effect of poverty was strong for conduct disorders; there was little evidence of an effect on outcomes like depression and social anxiety. Moreover, a similar pattern held for non-Indian participants in the study who did not receive stipends. Consistently with this, ex-poor families in both groups (that is, those who had moved out of poverty) reported an increase in time available for supervision (because both parents no longer needed to work full time). The

resulting picture is thus fairly coherent: it makes sense (in terms of our folk-psychological assumptions) that increased parental supervision should affect conduct disorders but there is less reason to expect it will affect outcomes like depression and this is what is found. (Note the role of specificity here.) It also makes sense that increased income that reduces the necessity of both parents working full (or more) time should increase parental supervision. Note, however, that this also seems to suggest that the increased income per se may not the active ingredient in reducing conduct disorder. If family income increases only because both parents work more, this will not contribute to a reduction in conduct disorder if the "mechanism" by which income affects disorder is as specified above. By contrast, an increase in income due to a stipend will work through the mechanism described above. This is thus an argument for the use of stipends/income supplementation.

13. Summary and Conclusion.

This paper has explored some of the strengths and limitations of various strategies for inferring causal relations involving mainly environmental variables in psychiatry. I have emphasized the advantages of design-based strategies such as RCTs and instrumental variables over more traditional strategies based on identifying and conditioning on possible confounders. However, these strategies can come with costs, including failures of generalizability and interpretability, as well as inattention to patient heterogeneity. The role of considerations like stability and specificity in controlling for possible confounders, as well as the benefits of triangulation strategies were also emphasized.

References

Baron, R. M.; Kenny, D. A. (1986). "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations". *Journal of Personality and Social Psychology.* **51** (6): 1173–118.

Blanchard, T, (2022) "Specificity of association in epidemiology" *Synthese* 200:482

Costello, E., Compton, S. Keeler, G. and Angold, A (2003) "Relationships between poverty and psychopathology: a natural experiment" *JAMA* 290(15):2023-9.

Davey Smith, G and Hemani, G. (2014) "Mendelian randomization: genetic anchors for causal inference in epidemiological studies" *Human Molecular Genetics* 23: Issue R1

Deaton, A. and Cartwright, N. (2018) "Understanding and misunderstanding randomized controlled trials" *Social Science and Medicine* 210: 2-21.

Esterling, K.., Brady, D. and Schwitzgebel, E. (n.d.). "The Necessity of Construct and External Validity for Generalized Causal Claims."

Freedland, K., Skala, J., Carney, R. (2009) "Treatment of Depression After Coronary Artery Bypass Surgery: A Randomized Controlled Trial" *Archives of General Psychiatry.* 66(4):387-396.

Hill, A. (1965) "The Environment and Disease: Association or Causation?" Proceedings of the Royal Society of Medicine **58** (5): 295–300.

Imbens, G., Angrist, J. (1994) "Identification and Estimation of Local Average Treatment Effects*" Econometrica.* 62 (2): 467-75.

Kendler, K. and Prescott, C. (2006) *Genes, Environment and Psychopathology*. New York: The Guilford Press.

Kendler, K. Ohlsson, H. Fagan, A. Lictenstein, P., Sundquist, J. and Sundquist, K. (2018) "Academic Achievement and Drug Abuse Risk Assessed Using Instrumental Variable Analysis and Co-relative Designs" *AMAPsychiatry* 75(11):1182-1188.

Rothman, K. J., & Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health, 95,* S144-150.

Rothwell,P. (1995)" Can overall results of clinical trials be applied to all patients?" *Lancet* 345, Issue 8965, m1616-1619.

Sellers R, Warne N, Rice F, Langley K, Maughan B, Pickles. A (2020) "Using a cross-cohort comparison design to test the role of maternal smoking in pregnancy in child mental health and learning: evidence from two UK cohorts born four decades apart." *International Journal of Epidemiology* 49:390–9.

Shrout, P. (2011) "Integrating Causal Analysis into Psychopathology Research" in Shrout, P., Keyes, K. and Kendler, K. (eds.) *Causality and Psychopathology*, pp 3-24.

Skoglund, C. , Chen, Q., D′ Onofrio, B. M., Lichtenstein, P., & Larsson, H. (2014). "Familial confounding of the association between maternal smoking during pregnancy and ADHD in offspring" . *Journal of Child Psychology and Psychiatry, 55(1), 61-68.*

Stuart, E., Schmid, I., Nguyen, T. Sarker, E., Pittman, A. Benke, K. Rudolph, K., Badillo-Goicoechea, E., Leoutsakos, J-M, (2021) "Assumptions Not Often Assessed or Satisfied in Published Mediation Analyses in Psychology and Psychiatry" *Epidemiologic Reviews* 43: 48–52,

Weiss, N. (2002). "Can the "Specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis?" *Epidemiology, 13,* 6–8.

Woodward, J. (2003) *Making Things Happen*. New York: Oxford University Press.