Commentary: Counterfactual causation and streetlamps: what is to be done?

James M Robins[1],* and Michael B Weissman[2]

[1]Harvard T.H. Chan School of Public Health, Boston, MA, USA and [2]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, USA

*Corresponding author. Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA. E-mail: robins@hsph.harvard.edu

Recent criticisms of the counterfactual causation programme in epidemiology have addressed ways in which the program not only inhibits consideration of the full ranges of evidence needed to infer causation but also excludes from consideration the effects on health of large structural and societal change, by implicitly restricting both the kinds of interventions considered and the type of questions asked. Here we respond to several of these criticisms, including ones with which we agree, ones with which we disagree and ones whose meaning we cannot discern.

Ill-defined counterfactuals

We the first address the issue of ill-defined counterfactuals as raised by Vandenbroucke *et al.*[1] One of us (J.M.R.) has been involved in a similar debate[2] before, defending (with Sander Greenland) the counterfactual approach against a claim by Phil Dawid[3] that counterfactuals are unscientific, even dangerous. It is striking that the Robins-Greenland defence against Dawid's accusation of their 'radicalism' is also a defence against the accusations of 'conservatism' by Vandenbroucke *et al.*[1] and Schwartz *et al.*[4] Robins-Greenland argued that although all counterfactuals are somewhat vague (think, obesity) and some even ill-defined, this vagueness only represents the fundamental fact that what we mean by the causal effect of a given exposure in an observational study is itself always imprecise and vague to a greater or less extent.

In fact, when a causal effect is so vague that there is no agreement about its meaning, one can only eliminate this vagueness to the extent possible by making more precise the hypothetical interventions one wishes to consider. Because counterfactual theory forces this problem of vaguely defined causal effects into the open, the argument that counterfactual theory is too restrictive is largely a 'shoot the messenger' response to this fundamental problem.

Indeed, although Vandenbroucke *et al.*[1] claim that 'limitation of epidemiology to one particular view of the nature of causality is problematic', a close reading shows that none of their arguments concerning the techniques best suited for estimating causal effects actually challenges in any way the counterfactual approach. Far from showing that the counterfactual 'view of the nature of causality is problematic', they have simply not addressed that issue (see Daniel *et al.*[5] and VanderWeele[6] in this issue of *IJE* for more detailed rejoinders to the Vandenbroucke *et al.*[1] paper).

A political critique

Another style of criticism of the counterfactual approach follows political lines. It is useful to begin by discussing for background a paper[4] in which the authors present a cogent criticism of the strongest version of the counterfactual causal inference approach. Schwartz *et al.*[4] argue that statistical data only yield firm conclusions about the consequences of alternative actions when one has either performed a randomized clinical trial (RCT) or successfully emulated a trial by applying the new causal methods to observational data, especially data in which one individual's treatment does not affect another's outcome. This limitation encourages a focus on individual-level 'treatment' effects, often incremental ones, rather than on the effects of large-scale social changes or movements.

One view of the Schwartz *et al.*[4] critique is that it applies only to the maximalist (perhaps straw-man) form of the counterfactual approach, However, we interpret the critique more broadly, bound up as it is with their pointed political objections. In fact, we agree with much of the Schwartz *et al.* critique. The streetlamps of the recent

causal methods illuminate some potential treatments much more directly than they illuminate others. That does not mean that the keys to fixing major social, political and public health problems are any more likely to be located under those lamps than they were before the lamps were turned on. (It should be noted that causal methods that relax the assumption that an individual's response is only affected by her or his own treatment are under rapid development, so illumination of additional classes of interventions are to be expected.[7,8])

Indeed, neither the recent causal methods nor other traditional epidemiological approaches have been of much help in answering questions of how to estimate the causal effects of large-scale social changes or movements or to identify interventions likely to bring them about. However, this does not reflect on these causal methods but on the difficulty and complexity of the questions, both conceptually and practically. The lack of even a minimal scientific consensus (even among progressives) as to the answers speaks to the current intractability of the questions. Some speculate that by combining experiments and observational analyses of massive social media, such as Facebook, an empirical science of social behaviour capable of engaging these questions may emerge. We shall see. Of course, there is no guarantee that such a science would be used for 'good' rather than 'ill'.

As a minor quibble, we disagree with the sharp distinction that Schwartz et al.[4].draw between the types of information needed by policy elites and by grass-roots activists. Often there is substantial overlap. For example, environmental activists need to know whether it's more important to focus on leaking residue from manufactured gas plants or on genetically modified soybeans. The same information about health effects is needed by policy makers.

The Krieger-Davey Smith paper:

With this background we turn the discussion to the paper by Krieger-Davey Smith (KDS) appearing in this issue of *IJE*. The KDS[9] paper criticizes the counterfactual approach from a variety of directions, some of which we believe add confusion rather than clarity to the discussion.

Consequentialism

KDS[9] write that they want a more 'robust causal inference' that goes beyond 'counterfactual and potential outcome reasoning'. With regard to the KDS[9] paper, although we are not followers of Lenin, it is important to face the question of 'what is to be done?' Although there are ways of answering that question which do not depend on the likely consequences of what is done, we think that most public

health discussions implicitly assume consequentialist ethics. In fact, KDS[9] introduce the importance of the epistemological argument by writing, 'The stakes, after all, are high: riding on the findings of epidemiological research are ...who and what is shaping population distributions of health, disease and well-being, within and across societies, and at what cost—and what benefit—to whom?' Later, in arguing against 'the current counterfactual framework', they argue about its misuse 'potentially causing harm'. It would be hard to formulate a more clearly consequentialist criterion for the choice of epistemology itself. Should we then adopt entirely different criteria for other choices? Despite obvious practical difficulties in making firm predictions, 'counterfactual and potential outcome reasoning' is by definition the way consequentialists answer the question 'What is to be done?'

Definitions of cause

KDS[9] raise a variety of distracting side issues and anecdotes, which we shall argue muddy the epistemological waters. KDS[9] describe their favoured approach as 'inference to the best explanation'. In the absence of a definition of 'best', that would just beg the question. They propose, following Lipton,[10] that 'best' largely consists of maximizing 'scope, precision, mechanism, unification and simplicity'. Maximizing precision, unification and simplicity is just standard scientific practice for any sort of proposition, regardless of whether it is specifically causal. In some cases sufficient relevant mechanisms are available to obviate the need for RCTs or their observational analogues to predict counterfactual outcomes. Predictions of the effects of merging black holes come to mind. Few, if any, advocates of counterfactual causal inference would reject this standard approach to science. In epidemiology, unfortunately, such precise and general law-like mechanisms are rarely found.

KDS argue that one should 'triangulate' a variety of lines of evidence to evaluate causal claims. We know of no one who would dispute that causal claims, like any other scientific claims, are most robustly evaluated by using diverse types of evidence. More problematically, KDS argue that one should not focus exclusively on any one of 'five families of "standard views" of causality'. These families give multiple meanings to various causal words. We are concerned that one might evaluate causal claims by 'triangulating' over multiple meanings. To evaluate the truth of some claim by taking some sort of composite of truth values of a variety of claims that might be expressed in similar words is fundamentally to mistake verbalisms for propositions. Conflating verbalisms with propositions is the opposite of a scientific approach that strives to disambiguate different propositions.

We first look at an admittedly extreme example, to make these philosophical abstractions more vivid. Say that Joe says 'Bill is gay'. That might mean, in traditional usage, 'Bill is lighthearted and cheerful'. It might mean in modern standard usage, 'Bill is homosexual'. It might mean in out-of-date vulgar teenage usage, 'Bill is yucky'. Would we evaluate the claim by 'triangulating' various lines of evidence for these three distinct meanings? Or would we try to disambiguate them and perhaps come to entirely distinct conclusions about their truth values, if any? Likewise, if there are five families of different meanings for the claim 'A causes B', wouldn't it make most sense to try to clarify those different meanings and evaluate the objectively meaningful ones separately? In a recent paper 'Does water kill?', Miguel Hernan argues along similar lines. Hernan makes clear that the meaning of this question is so vague as to be meaningless without further specifying the intervention under consideration

The specific examples cited by KDS to support the alleged need for triangulating causal claims are puzzling. Some are simply matters of scientific surprises occurring, for example the anatomical dispersal of olfactory receptors or the multiple physiological effects of estrogen. New scientific findings can cause us to change our minds about all sorts of assertions, including ones about causal relations. There are families of assertions (e.g. some religious claims) that are held to be immune to evidence, but we do not think that such assertions have a place in epidemiology. As Keynes is alleged to have remarked, 'When my information changes, I alter my conclusions. What do you do, sir?' So the possibility that counterfactual causal claims might be incorrect and require revision in the light of new evidence simply means that they are part of science, not that they are particularly narrow or weak.

Examples of unavoidable issues in the definition of causal effects

Indeed, it is often the case that a counterfactual that epidemiologists and other scientists consider well-defined at one point in time is later understood to have been rather vaguely defined, in the light of new scientific findings. As an example, there was rather substantial agreement in the 1960s that the causal effect of serum cholesterol on coronary heart disease (CHD) was sufficiently well-defined that the effect of lowering serum cholesterol on CHD incidence in subjects with hypercholesterolaemia might possibly be quantitatively predicted on the basis of the Framingham data. Today, in light of our increased knowledge of the differing consequences of high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol for CHD risk, that agreement seems naïve. Fortunately, that time-limited

agreement was sufficient to motivate RCTs of cholesterol-lowering agents that indeed proved beneficial, even though a controversy remains as to whether other, non-cholesterol related pathways contribute to the benefit.

There have been attempts to define the causal effect of a deterministic function (e.g. total cholesterol = HDL + LDL) of a multivariate or composite exposure (e.g. HDL and LDL), each component of which is presently thought by some to possess reasonably well-defined counter-factualsl.[12,13,14] The approach has been based on representative regimes as defined in Taubman *et al*.[12] The representative regime that sets serum cholesterol to 100 is defined to be an intervention in which, for every subject and time, HDL and LDL are sampled at random from their observed distribution conditional on both their sum equalling 100 and possibly additional covariates. However, this approach is far from fully resolving the issues raised by ever-increasing scientific knowledge. First, the cholesterol effect will differ depending on the additional covariates that the regimen conditions on. Second, the effect of cholesterol will differ over time, even if the effects of HDL and LDL do not, since the joint distribution of HDL and LDL in 2016 differs from that in the 1960s. Third, such a regimen often could not be implemented in reality.

Next suppose one wishes to estimate the 'causal effect of body mass index (BMI)' on longevity. We suspect that most epidemiologists implicitly interpret the 'causal effect of BMI' as the effect of an atomic intervention in which fat cell mass is directly increased or decreased. Under this interpretation, recent exercise should be viewed as a time-dependent confounding factor, likely affected by earlier BMI. Hence g-methods should be used to estimate BMI's effect. We would argue that this is the case even though the meaning of the 'causal effect of BMI' on mortality (and thus the associated counterfactual) remains ill-defined, since two nearly atomic interventions, such as weight loss stomach surgery and liposuction, may have differing causal effects, even when they produce identical changes in BMI. The point is that no matter which of these causal effects one is interested in, it is likely that the population quantity (i.e. estimand) being estimated by the g-formula will likely be closer to the causal effect of interest than the population quantity being estimated by a method that fails to appropriately adjust for time-dependent confounding by exercise. Of course, for policy purposes, estimating the effect of a well-defined implementable intervention would be preferred to estimating 'the effect of BMI'.

Kds's examples

Some KDS examples seem entirely orthogonal to the question at hand. (The commentary by Daniel *et al*. addresses

the following examples in greater detail.) One KDS example concerns the causes of pellagra. The variety of proposed causes, each leading to different implications for the results of deliberate or accidental interventions (e.g. on diet), seems like a textbook case to illustrate standard counterfactual causal methods, complete with different DAGs. We are at a loss to see how it supports the KDS argument. Similar comments apply to the example of Semmelweis and child-bed fever. A third KDS example is the birthweight mortality 'paradox'. As they point out it, has been described non-paradoxically within the DAG framework. As they also point out, the structure of the DAG explanation does not in itself specify which particular genetic or environmental factors are mainly responsible for the high mortality rate in low-birthweight infants of non-smokers. It's true that, in general, stating that a correlation may be due to 'conditioning on a common effect of independent causes' [i.e. collider bias] is much less informative than specifying what those independent causes are. Is there someone to whom that needed to be pointed out?

One might then wonder why KDS felt that these examples were needed. We of course cannot know, but we do have a conjecture. To have a methods paper accepted for publication in a leading epidemiology journal requires that the paper either contain, summarize or apply new methods. Since the theory and application of formal counterfactual causal methods are undergoing rapid and novel development, it is natural that many such papers are being published. In contrast, the scientific methods exemplified by the child-bed fever and pellagra examples are so well-known that they are included in the curriculum of all training programmes in epidemiology. As a consequence, the current methodological research literature naturally appears highly skewed towards papers on counterfactual causality which, we conjecture, motivated at least in part the KDS and Vandenbroucke critiques. (The flip side is that many departments of epidemiology are yet to include courses in counterfactual-based causal methods in their curriculum.)

Race

A central objection of KDS is to the treatment of 'race' in the counterfactual causation literature. Briefly, even if we could set aside the issue of whether race is a well-defined category, there would still be little agreement as to the meaning of the question 'does race have a causal effect on cognitive function at age 50?' Both a progressive social activist and a White supremacist might answer 'yes' to the question. Is their agreement useful or meaningful? The problem is, of course, that racial effects include genetic effects, effects of ancestral social environment, effects of ongoing political disenfranchisement and material inequality,

effects of social reactions to perceived race and so on. Without further disambiguation, to ask whether race is a cause is only slightly more useful than our asking earlier whether Bill is gay. Once disambiguated, one can even quantify some of these effects using RCTs, for example by randomly assigning racially distinctive names to case histories or rental applications.

KDS's objection to the treatment of race in the potential outcome literature is 2-fold, with some inconsistency between the folds. For KDS, on the one hand 'race' does not exist; on the other hand it cannot be dismissed as a cause. After conceding that there can be problems in invoking a non-existent trait as a cause, KDS say that really it's 'racism' that's being considered as a cause. This view is quite consistent with the counterfactual framework, since modifications of racism (and blind experiments to test some of its effects) are feasible, and to some extent already done. With regard to the direct biological effects of those genes present at different frequencies in different socially-defined racial groups, one of the authors of KDS has written extensively on the possibility of inferring effects of some genes using the somewhat random reshuffling of genes in the population.

It would be interesting to know the authors' views of studies of medications, for example antihypertensives, that search for 'qualitative interaction' (technically, effect modification) by (usually) self-identified race. Doesn't the National Institutes of Health's (NIH) insistence that minorities be well-represented in any study reflect a desire to be able to empirically screen for such qualitative effects? Obviously, if after adjustment for other factors, such an interaction is found to be robust, one needs to investigate alternative causal hypotheses. Is the cause of the interaction socially determined rather than genetic? If genetic, which genes are responsible, as any causal variants are likely present in other (self-identified) ethnic groups? Furthermore, even if genetic, an interaction with the downstream effects of racial discrimination might be required for the effect to be present. Until genomic sequencing becomes standard, it may take months, perhaps years, to investigate these alternate hypotheses. Until then, it seems prudent to withhold the treatment from those ethnic groups in whom it has been found harmful.

In summary, the work of KDS ends up arguing that there are a variety of causal effects of aspects of racial differences that can be tested without attempting to reify the concept of race. In particular, the various roles that social environment (e.g. racism) plays in mediating those events are clearly testable using a variety of study designs. Thus we are puzzled as to what the real objection of KDS is. We return to this issue below.

Suburbs and climate change

KDS conclude with a discussion of a charming 1957 passage from Morris speculating on the effects of broad social changes: increased going-to-work of married women, increased suburbanization, reduced physical activity etc. These questions are important in deciding what is to be done, and evidently cannot be fully answered by real-world RCTs. As a consequence, we must estimate the causal effect of, say, suburbanization under considerable uncertainty. One approach is to try to inform our decisions by putting together and extrapolating from different, indirect, types of evidence, for example changes of cortisol levels associated with automobile commutes. Such indirect evidence should, when possible, be evaluated in randomized experiments or, barring that, using modern counterfactual causal methods. In addition, since rates of urbanization vary considerably among cities, by treating the inhabitants of each city as a cluster we can try to estimate the effect of urbanization using current causal methods. Depending on the specifics, adequate control of confounding is often, but not always, more problematic than in typical studies of independent individuals. Ultimately all forms of relevant evidence must be weighted by reliability and relevance and synthesized either informally or within a formal Bayesian analysis. In any case, the form of the question: 'How will our lives differ if we build suburbs versus dense walkable cities?' remains firmly within the counterfactual framework, even though the interventions under consideration remain somewhat vague, without further attempts to make them more specific.

It is perhaps instructive to consider a canonical case where an RCT is infeasible even in principle, yet a synthesis sufficiently compelling to result in a scientific consensus has been reached, although there remain a few who disagree. Specifically, the effect of greenhouse gases on climate change cannot be evaluated by the maximal counterfactual programme of performing specific RCTs to test alternative treatments. We cannot try emitting different amounts of CO_2 and CH_4 on a variety of randomly-assigned Earth-like planets to see which ones are most severely affected. Nevertheless, we can put together and extrapolate from different types of evidence, using standard scientific methods. For starters, we know that adding greenhouse gases to an atmosphere slows the outflow of infrared radiation. Then, using a variety of approximate climate models based on such well-known physical laws and constrained by a plethora of diverse observational data (from tree-ring thicknesses to isotopic content of ancient ice to satellite measurements of atmospheric temperatures at a range of altitudes), it is possible to obtain fairly reliable estimates of what the different outcomes will be for different inputs of these gases. Unfortunately, the phenomena studied in epidemiology are more remote from such reliable laws.

Politics and science

Finally, we think that in their discussion of race and social change, KDS papers over real tensions that confront any progressive population scientist. Specifically, although we have argued the advantages of a formal counterfactual approach, we recognize that such an approach draws attention to the limits of our knowledge and to uncertainty. Motivating popular action on contested social questions may require downplaying both limits and uncertainties, if one wishes to avoid playing into the hands of a powerful opposition who benefit from the status quo. Here we have only considered the role of epidemiologists in providing information for consequentialist decisions, not in helping motivate actions. It would be interesting to hear the authors' views on this issue.

Summary

The exchanges for and against the counterfactual approach to causation to this point appear to exhibit much mutual misunderstanding about what different players advocate, leading to many 'straw-man' complaints. Perhaps it would best to give a very brief credo, rather than further arguments.

i. The meaning of the word 'cause' that is relevant to consequentialist decisions is the counterfactual meaning. Other meanings or conceptualizations may, however, be useful for other purposes.

ii. Estimates of relatively well-defined causal effects can be made by the full panoply of scientific techniques, including inferences based on known laws, without any hope of conducting an RCT.

iii. In areas where the laws are unclear, unknown, imprecise or may not exist, RCTs and RCT-like analyses of observational data can provide more or less reliable, relatively assumption-free estimates (in the case of randomized trials) of relatively well-defined causal effects that are unavailable by other techniques.

iv. The causal effect of a given exposure in an observational study is always vague to a greater or lesser extent. Further specifying the intervention under consideration will reduce the degree of vagueness.

v. The ability to obtain reliable answers to some questions with the new causal methods does nothing to make other questions less important. That's true both of 'hard science' questions already answerable by mechanistic physical law and of large social questions not currently answerable with confidence.

References

1. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int J Epidemiol* 2016;**45**:1776–86.
2. Robins JM, Greenland S. Comment on 'Causal inference without counterfactuals.' *J Am Stat Assoc* 2000;**95**:477-82.
3. Dawid AP. Causal inference without counterfactuals. *J Am Stat Assoc* 2000;**95**:407-24.
4. Schwartz, S, Prins SJ, Campbell UB, Gatto NM. Is the 'well-defined intervention assumption' politically conservative?, *Soc Sci Med* 2015; http://dx.doi.org/10.1016/j.socscimed.2015.10.054
5. Daniel RM, De Stavola BL, Vansteelandt S. The formal approach to quantitative causal inference in epidemiology: misguided or misrepresented? *Int J Epidemiol* 2016;**45**:1817–29.
6. VanderWeele TJ. On causes, causal inference, and potential outcomes. *Int J Epidemiol* 2016;**45**:1809–16.
7. VanderWeele TJ, Tchetgen Tchetgen EJ, Halloran ME. Interference and sensitivity analysis. *Stat Sci* 2014;**29**:687–706.
8. Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc* 2008;**103**:832–42.
9. Krieger N, Davey Smith G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol* 2016;**45**:1787–808.
10. Lipton P. *Inference to the Best Explanation*. 2nd edn. London: Routledge, 2004.
11. Hernán, MA. Does water kill? Causal inferences anchored to target trials or how to make less casual causal inferences. *Ann Epidemiol*, in press.
12. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Alternative approaches to estimating the effects of hypothetical interventions. *Joint Statistical Meetings Proceedings Health Policy Statistics Section, 3-7 August 2008, Denver, CO*. Alexandria, VA: American Statistical Association, 2008.
13. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. *Epidemiology* 2011;**22**:368-77.
14. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. *J Causal Inference* 2013;**1**:1-20.

Commentary: DAGs and the restricted potential outcomes approach are tools, not theories of causation

Tony Blakely,[1,3]* John Lynch[2] and Rebecca Bentley[3]

[1]Department of Public Health, University of Otago, Wellington, New Zealand, [2]School of Population Health, University of Adelaide, Adelaide, SA, Australia and [3]School of Population and Global Health, University of Melbourne, Parkville, VIC, Australia

*Corresponding author. Department of Public Health, University of Otago, Wellington, New Zealand. E-mail: tony.blakely@otago.ac.nz

Accepted 20 July 2016

There must be few epidemiologists who do not use directed acyclic graphs (DAGs) in their teaching and research as tools. And this is the key point; they are useful tools, not theories of causation.

Perhaps not as widely used as DAGs is the potential outcomes approach (POA), a bedrock of many causal inference methods, and to those users of POA probably as useful (if not more so) than DAGs. For example, VanderWeele has led the way in using the POA (and the umbrella counterfactual way of thinking) to advance methods for mediation analysis.[1] Such methodological development is welcomed, albeit perhaps pushing ahead on structural relationships of variables while (largely) ignoring the likely important issues of measurement error and underlying data quality.[2,3] So it may